Explaining register and sociolinguistic variation in the lexicon: Corpus studies on Dutch

Published by

LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6006
e-mail: lot@uu.nl
http://www.lotschool.nl

ISBN: 978-94-6093-090-4
NUR: 616

Explaining register and sociolinguistic variation in the lexicon: Corpus studies on Dutch

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de Rector Magnificus
prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op
maandag 15 oktober 2012
om 13.30 uur precies

door

Karen Keune

geboren 4 september 1979
te Nijmegen

Promotoren:	Prof. dr. R.W.N.M. van Hout Prof. dr. R. H. Baayen (Eberhard Karls University, Germany)
Manuscriptcommissie:	Mevr. prof. dr. M. van Mulken (voorzitter) Prof. dr. J. Nerbonne (Rijksuniversiteit Groningen) Mevr. prof. dr. S. Tagliamonte (University of Toronto, Canada)

voor Stan en Tijn

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Texts consist of a broad and varying spectrum of lexical elements. There is variation in the number of nouns, adjectives or verbs (word classes), variation in the amount and share of complex words, variation in the degree in which highly frequent and function words occur, and variation in the number of unique words in a text. How can we analyze such patterns of lexical variation?

The choice and use of words, word groups, and word classes appears to be highly dependent on the audience, the function and the medium of the message. To cover the role of different language situations or contexts of language use and the purpose of a text, the term 'register' is used (Reid, 1956; Halliday, 1964; Biber, 1988, 1995; Biber and Conrad, 2009). This term was introduced to distinguish between language variation according to the *user* (regional and social variation, resulting in differences between speakers) and according to the *use* (register variation). Halliday and Matthiessen (2004) define registers as 'ways of using the language'. Register includes the functional varieties of language, in principle available to all speakers. This contrasts with defining regional and social varieties, which basically relate to the characteristics of the speaker or user. The concept of register differs from the concept of genre, in that register captures the lexico-grammatical features of a text, while genre captures the context itself in which a text is produced (van Dijk, 2009).

The words, lexical elements, used for instance in a newspaper, will differ from the lexical elements used in an informal email message. While a newspaper will probably have a more informational purpose, it is likely that an email message will in general contain more lexical elements that reflect personal involvement. This results into two different registers within written language. Within the register of newspaper articles, again different registers can be distinguished. The register 'quality newspaper'[1] is distinct from the register

[1] We use the term 'quality newspaper' for a newspaper aiming at a highly educated readership.

'tabloid'[2], for instance, aiming at a different readership, and it is to be expected that they adapt the language, including the choice of the lexical elements, accordingly. The register differences between a quality newspaper and a tabloid are probably smaller than the register differences between a newspaper and an email message. Register differences can of course also be found in spoken language. A prepared presentation (monologue) will differ from an unprepared, spontaneous telephone conversation (dialogue).

Lexical differences between registers can be illustrated by the two following sentences taken from two different registers:

1. Watercondensatie doet de temperatuur immers stijgen en daardoor is ook nog een gradiënt te verwachten in het grensgebied tussen wolk en heldere lucht.

   Water condensation, in fact, makes the temperature rise and as a result a gradient is expected too in the border area between cloud and clear sky.

2. 'een uh soort *a v*a soortement verkoper managerachtig en uh die moet zeg maar de producten die wij dan maken of in ieder geval kunnen moet hij proberen aan de man te brengen.'

   *'A uh kind*a*[3] *v*a kind of salesman, manager like and uh that one has, so to say the products we make or at least are able to, he has to try to sell them.'*

It will be no surprise that sentence 1 comes from an informative, formal written text. The sentence is cited from the Dutch quality newspaper 'NRC Handelsblad'. It is also directly visible, that sentence 2 is an orthographic transcription of a speech fragment. It comes from a spontaneous, informal dialogue from the Spoken Dutch Corpus (Oostdijk, 2002).

The first sentence is grammatically correct, fluent and dense in information. The second sentence contains hesitations, recaptures, and complex words are thought of during speaking: *uh kind*a v*a* becomes *soortement*.

When we focus on the lexical elements of the two fragments, we see that Sentence 1 contains 23 words (word tokens), 22 of these words being unique (word types). The word *en* ('and') occurs twice. Furthermore, it contains ten words belonging to the 80 most frequent words (also called most common words) of the newspaper corpus it occurs in, and it contains one word ending in a derivational affix: *watercondensatie*. It contains six nouns, one adjective and four verbs.

Sentence 2 contains 33 word tokens and 29 word types: four words occur twice. It contains 18 words that are among the 80 words that are used most frequently in the Spoken Dutch Corpus. It contains two derivational word forms: *soortement* (soort = *kind,* -ement = *suffix: of*) , and *managerachtig* (*manager like = like a manger;* -achtig = *-like*), three nouns, one adjective, and six verbs.

[2]In Chapter 2 and Chapter 3 we use the term 'national newspaper' instead of 'tabloid' to refer to this newspaper register.

[3]Hesitations in the CGN are marked with a *. See also Oostdijk, 2002.

1.1 Corpora

These two sentences only give us a first glance into the kind of lexical differences that one finds across registers. We need more sentences to draw more robust conclusions. To obtain insight in the variable use of lexical items across registers, we need to explore major text corpora. For the present research, we explored two major Dutch corpora. The first is the CONDIV corpus (Grondelaers et al., 2000), a corpus that contains much data from internet sources (Usenet and Internet Relay Chat), and many newspaper articles from seven different newspapers. For both the Netherlands and Flanders, this corpus contains a quality newspaper, a national tabloid, and a regional newspaper. This makes it possible to not only compare and contrast country (the Netherlands versus Flanders) but to compare different newspaper registers too. In addition, the corpus contains newspaper articles from 1958, 1978, and 1998, which gives the opportunity to explore the diachronic dimension. The complete corpus comprises approximately 47.4 million words, of which approximately 17.6 million words come from newspapers.

The second corpus we used is the Spoken Dutch Corpus (CGN) (Oostdijk, 2002). This corpus contains approximately 8.9 million words from speech fragments of Dutch and Flemish adults. Within the corpus, subcorpora are formed distinguishing various speech registers. The corpus contains private speech (unscripted conversations and telephone dialogues: 4.7 million words), public speech (3.4 million words), and read aloud speech from the library of the blind (0.9 million words). The public speech part can be split up into dialogues (for instance debates, meetings, and interviews: 2.3 million words) and monologues (news, reportages, and commentaries (all broadcast), reviews, ceremonious speeches, and lectures: 1.1 million words).

Speaker information such as country, gender[4], education level, and age is available, which makes it possible to study the influence of these sociolinguistic factors on the use of lexical elements.

Both the CONDIV corpus and the Spoken Dutch Corpus are widely used in linguistic research. The CONDIV corpus is predominantly used in sociolinguistic research. Grondelaers et al. (2001), for instance, explored whether Netherlandic Dutch and Belgian Dutch converged or diverged between 1958 and 1998 by comparing clothing and soccer terms (see also Geeraerts et al., 1999) and prepositions, and investigated whether the distance between formal and informal language is larger in Belgian Dutch than in Netherlandic Dutch. They found an obvious convergence in the language in the above-mentioned period of time, and also showed that the distance between formal and informal language was larger in Belgian Dutch.

The Spoken Dutch Corpus has been used in different kinds of corpus research. All material is orthographically transcribed, lemmatized, POS-tagged,

[4]In Chapter 2 and Chapter 3 we use 'sex' instead of 'gender' to refer to the difference between men and women.

and linked to the speech signal. A selection of one million words was syntactically analyzed, and phonetically transcribed. The phonetic transcriptions make the corpus suitable for phonetic research. Van de Ven et al. (submitted) investigated to what extent listeners can use context to process low-predictable words in natural spontaneous speech. They found that listeners used both the context preceding and following the low predictable word. Schuppler (2011) investigated the many acoustic reductions in spontaneous speech. Plevoets (2008) explored the situational, regional, and social distribution of a large number of morphosyntactic elements in spoken Belgian Dutch in order to investigate the role of 'Tussentaal' ('inter language' between the standard language and dialects). Quené (2008) used interviews with secondary school teachers to investigate variation in speech rate. Phrase length, and the speaker's home community appeared to be the most important predictors of speech tempo. In the Netherlands the speaking style is faster, and less varied than in Flanders.

1.2 Stylistics and sociolinguistics

The lexicon is frequently investigated in corpus research with the aim of charting registers, especially in stylistic corpus research. Halliday (1978) introduced an abstract distinction between register and dialect. Dialect refers to who the speaker is, in a regional and social sense, while register refers to what use is being made of the language. According to Halliday and Hasan (1976) there are three aspects of language use that determine register: field (activity in which the text-producer is participating), tenor (social relation between producer and consumer), and mode (medium by which the text is produced). The term register is frequently related to the degree of formality of language. However, according to Halliday (1978), it may be better to use the term 'tenor' instead. Another popular term to refer to the degree of formality of language is 'style'. Joos (1961), for instance, described five styles (levels of formality) in spoken English situated on a linear scale of 'formality', labelled 'frozen', 'formal', 'consultative', 'casual', and 'intimate'.

Biber has carried out probably the most extensive studies on register variation in corpus linguistics (Biber, 1988, 1995; Biber and Conrad, 2009). These studies analyzed differences among written and spoken registers of the English language on the basis of a broad range of linguistic characteristics. Written texts turn out to be more 'informational', while spoken texts are more 'involved'. The linguistic characteristics typical of informational texts were, among others, long words, more prepositions, and the use of a larger number of different words, whereas present tense verbs, and the use of the word 'you' were most typical of involved production.

Burrows (1992a,b, 1993a,b) explored variation across regions, but also across individual writers in literary studies. He developed a stylometric technique to identify individual language users on the basis of their use of the most common words. These words typically include function words as well as some frequent

adverbs. Baayen (1996) concluded that differences in the use of these most common words tend to represent differences in syntactic habits. In the field of corpus linguistics, the impact on the lexicon of external, sociolinguistic variables, such as gender, education level, age, and also region, remained generally underexposed. Most corpus studies on lexical variation in the field of stylistics do not include regional and social variables. The studies that do include sociolinguistic speaker characteristics are mainly studies on written data. Newman et al. (2008) found systematic gender differences in 14,000 text samples for a series of lexical properties related to word classes and semantic fields. Argamon et al. (2003) revealed gender differences in the use of pronouns and certain types of noun modifiers in formal genres from the British National Corpus. There is, however, some research on global lexical variation in speech. Härnqvist et al. (2003) investigated 415 Swedish interviews and found differences in vocabulary richness and Part of Speech for gender and education level. Van Gijsel (2007) revealed that the speaker's country (the Netherlands versus Flanders), gender and age are, next to register, influential factors predicting lexical richness.

In the field of sociolinguistics, variables referring to the *user* rather than to the *use* are widely investigated. These *user* variables include regional variables (comparing speech communities) and social variables (such as age, gender, class, and education level). Sociolinguists studying variation, however, focus on well-defined separate linguistic variables, with a biased, but keen interest in phonological and morpho-syntactic linguistic variables. Defining a linguistic variable means that the variants (the forms) belonging to the same linguistic phenomenon are being investigated. Jacobi (2009) used the Spoken Dutch Corpus to explore the role of speaker characteristics on the pronunciation of the six long vowels and diphthongs in Dutch, with a special focus on the (*ij*), and found that the speaker's education level, age, and gender were all significant predictors. Van Bergen, Stoop, Vogels and de Hoop (2011) investigated the occurrence of the pronoun *hun* ('them') in the Spoken Dutch Corpus as the new competing variant of the standard variants *ze/zij* ('they') in subject position. The occurrences were not so frequent that they could establish the impact of social variables, but they traced relevant linguistic conditions.

Variationist studies in sociolinguistics tend to define their linguistic variables in terms of varying forms covered by the same meaning. However relevant this discussion is, the differences between a register and stylistic approach on the one hand and a sociolinguistic one on the other, are sharply delimited with respect to the language phenomena investigated. Sociolinguists investigate smaller pieces of specified linguistic elements, perhaps in a more global perspective of connecting all these elements in terms of varieties, but the elements remain the building blocks. The stylistic or register perspective principally defines lexical patterns in a global way by distinguishing word classes and word groups, larger semantic fields and lexical indices, for instance of lexical diversity and lexical density.

Another relevant distinction is the definition of style, which is still an un-

derdeveloped area in variationist sociolinguistics. Although one of the most impressive innovations of Labov (1966) was the application of the concept of style in explaining variation, style is still mainly interpreted as the amount of attention paid to speech (cf. Eckert and Rickford, 2011).

1.3 Research question

In this dissertation we aim to connect the fields of stylistics and sociolinguistics in studying lexical variation. The main goal of this dissertation is to expand our knowledge about the effect of register and the user variables country, gender, education level and age on variation of patterns of lexical distribution between and within written and spoken Dutch. We investigated newspapers articles from the CONDIV corpus and speech from the Spoken Dutch Corpus (CGN). Our aim is to explain register and sociolinguistic variation patterns in the Dutch lexicon.

We have investigated variation patterns of several components of lexicon variation: derivational productivity of the Dutch suffix *-lijk* (*-like*), general derivational productivity, and its relationship to the role of individual derivational suffixes, overall lexical productivity, and the share of the number of most common words and word classes (nouns, adjectives and verbs).

1.4 Lexical variation

How can we best measure lexical variation and patterns of lexical variation? As shown by the two example sentences in the beginning, a text is built up by different kinds of lexical elements. How can we define and handle more global lexical patterns, trying to get at a more abstract level of tracing text characteristics? An appealing example is the concept of the lexical density of a text, as measured by the relative share of content words occurring in that text. Johansson (2008) compared this measure to the measure of lexical diversity in a developmental perspective. She found a more noticeable developmental trend for lexical diversity than for lexical density. Another example is the relative share of word classes. Heylighen and Dewaele (2002) show that the relative share of word classes does not have the same, constant distribution over different registers. As the register of a text becomes more formal, the number of nouns, adjectives, prepositions and articles increase, and, as the contextuality of a discourse increases, the number of pronouns, verbs, adverbs, and interjections increases. The relative share of the most common words (words with the highest frequency) that appear in a text, has proven to be a good measure in authorship attribution (Burrows, 1992a, 1993a) and in pointing out differences in syntactic habits (Baayen, 1996).

Lexical variation is often measured by determining the lexical diversity of a text (or corpus). An extensively studied and frequently applied measure for lexical diversity is the Type-Token Ratio (TTR: Tweedie and Baayen, 1998;

Arnaud, 1984; Richards, 1987). This measure calculates the lexical diversity (or lexical richness) of a text by dividing the number of unique word forms (types) by the total number of words (tokens) in that text. Another measure of lexical diversity is the growth rate of the vocabulary size of a text measured by counting the number of hapax legomena (i.e. words occurring only once) and dividing this number by the total number of words in the text Baayen (2009). If a text with a high growth rate of the vocabulary size would be extended, it would give many new words, as the lexicon would still contain many unused words.

New words can be the product of a range of different word formation processes as well, such as word compounding and composition, word borrowing, and derivational word formation. Derivational productivity is probably, next to word compounding, the most interesting word formation process to investigate, since affixes can be highly productive (Baayen (2009) gives an overview of ways to measure derivational productivity). We included measures of derivational productivity in analyzing the CONDIV corpus and the Spoken Dutch Corpus.

1.5 Method of analysis

To measure register differences and related sociolinguistic patterns, we subdivided our newspaper and speech corpus into subcorpora distinguishing specific registers and regional (country) and social (gender, age education) variables.

In the Spoken Dutch Corpus the sizes of the resulting subcorpora turn out to vary enormously. For old Flemish female speakers with a lower education level, for instance, there were only a couple of thousand words available, while for young Dutch female speakers with a high education level hundreds of thousands of words were sampled.

As a consequence we have to evaluate different statistical methods to find the ones that can handle our type of data best. We will show that a Principal Component Analysis (PCA) is a fruitful technique to obtain a first global overview of the patterns in our data. To analyze the patterns more accurately, to find word or affix-specific information, and to include interactions terms between the predictors, we fit several types of regression analyses to our data. What measure fit best is, among others, dependent on whether the data is subdivided in samples equal in size or in subcorpora highly varying in size, on whether the data contains random or only fixed variables, and on whether the data is normally distributed or not.

We will also pay attention to methods of sampling the data, as we claim that repeated random sampling of corpus data might produce more robust figures than treating the full corpus data as the sample. This conclusion applies of course to the type of lexical measures we used.

1.6 Outline

In Chapter 2 we investigate variation in the frequency of use (written and spoken) and the degree of acoustic reduction (spoken) of 32 Dutch words ending in the suffix *-lijk* ('-like') to obtain a first impression of the presence of systematic global variation of this suffix in the CONDIV corpus and the Spoken Dutch Corpus. It is hardly productive anymore (Van Marle, 1988), and many high-frequency forms are no longer semantically compositional. The word *natuurlijk* ('nature like') for instance, has often lost its original meaning and is now used as a function word meaning 'of course'. The loss of the original semantic meaning seems to make it possible for speakers to pronounce such words in a highly reduced form. For instance *natuurlijk* is frequently pronounced as *tuuk*. We first investigate systematic variance patterns in the frequency of use of the suffix *-lijk* in written Dutch as a function country (the Netherlands versus Flanders) and newspaper register, and in spoken Dutch as a function of country, gender, and education level. To test whether the variation patterns we find are specific for words ending in *-lijk* or whether these patterns are reflected in other aspects of lexis and grammar, we create a lexical benchmark. We use the well-established stylometric technique developed by Burrows (1992a, 1993b) in which the most common words are explored to reveal variation patterns. Second, to reveal the influence of the speaker's country and social characteristics on the degree of reduction in the pronunciation of words ending in *-lijk* occurring in spontaneous speech, we make phonetic transcriptions of all occurrences of the 24 *-lijk* words that occurred frequently enough in the corpus to take them in consideration for further study. Only 14 of these words have a reduced form. In the analyses we distinguish three categories of reduction and take into account two statistical measures: the effects of the relative frequencies of the *-lijk* words in the sub-corpus it occurs in, and the mutual information between the word itself and the preceding word, which estimates the predictability of a word given the preceding word in the sentence. Furthermore we code whether the word occurred in the final position of a sentence or not.

In Chapter 3 we explore the potential productivity or the expected growth rate of derivational word forms in the Spoken Dutch Corpus (CGN), using 72 different Dutch affixes. For each affix we count the number of hapax legomena (i.e. the words that occur only once) having that specific affix in the spontaneous speech fragments. We distinguish 24 sub-corpora, as defined by the speaker's country, gender, education level, and age, and we investigate the distribution of the hapax legomena over the sub-corpora. The large number of cells with zero counts and the substantial variation in the sizes of the sub-corpora underlying the cell counts pose a particular challenge for the statistical analyses. We investigate the fit of three different statistical models. An ordinary least squares linear model with the transformed proportions of hapax legomena as dependent variable, a linear mixed effects model with affix as random effect and the transformed proportions as the dependent variable, and a gen-

eralized linear model with a binomial link, considering the hapax legomena as successes and all remaining words as failures. We investigate whether there is a global sociolinguistic variation pattern, and whether there are affix-specific differences.

In Chapter 4 we investigate the effects of register and sociolinguistic (country, gender, education level, age) variables on derivational and lexical variation in both written and spoken Dutch. We include lexical productivity in this study to clarify whether the variation patterns in derivational productivity are reflected by variation patterns in lexical productivity, or whether these two measures indicate separate and independent parts of the lexicon. Next to speech from the Spoken Dutch Corpus, we include written texts from the CONDIV newspaper corpus. This enables us to investigate the differences in affix productivity between spoken and written Dutch. We first compare derivational and lexical productivity in the main sub-corpora of newspaper Dutch (quality, national/tabloid, and regional newspapers) and spoken Dutch (formal monologue – public speech, dialogue – public speech, dialogue – private speech) and to compare the degree of productivity in the main registers of written and spoken Dutch. Next, we explore whether the resemblance, or difference in productivity of the individual affixes between written and spoken Dutch is a global effect or an effect carried by affix-specific differences. We use the generalized linear model to investigate the variation patterns within written and spoken Dutch in more detail. For written Dutch, we include country, and for spoken Dutch we include the speaker's country, gender, education level, and age to our analyses to analyze derivational and lexical variation in more detail.

In Chapter 5 we study sociolinguistic patterns in general lexical characteristics of Dutch spontaneous speech form the Spoken Dutch Corpus (CGN), to investigate the relevance of including such measures in the domain of sociolinguistic variation studies. As in the previous chapters, we include the speaker's country, gender, education level, and age. We include three types of lexical measures, namely lexical diversity (measured by the number of types, and number of hapax legomena), lexical density (measured by counting the number of nouns, adjectives, and verbs), and lexical communality (measured by counting the number of most common words). We stratify the corpus by applying our four speaker variables mentioned above. This resulted in 24 strata or sub-corpora strongly varying in size. We decided to work with samples from these sub-corpora because of the text length dependency of the measures of lexical diversity and because of the possible effects of topic dependency. We show the positive effects of random sampling by its strong effect on the reduction of topicality by comparing context samples (drawing connected text parts from the sub-corpus) with random samples (drawing words randomly, without replacement, from the whole sub-corpus). We use Principal Components Analysis to obtain an overview of the global variation patterns in the multivariate distribution of the six lexical measures and the four speaker characteristics. Next we apply linear regression modelling (separately for each of six lexical mea-

sures) to investigate the patterns in more detail, including the way the speaker characteristics interact. Finally, to obtain a better understanding of variation in lexical diversity, lexical density and the most common words, we investigate the contribution of individual words to the effects. We distinguish word-specific from global effects. If an effect is due to one specific word or a small group of words, it is word-specific. If an effect is the consequence of all words involved, affecting all lexical items, it is global.

In Chapter 6, we present the conclusions of the chapters and we will draw more general conclusions in relation to lexical measures, the role of register, and the impact of the regional and social variables. In addition, we will reflect on the methodology applied and the type of statistical analyses used. In the discussion section we will raise the question of the match we tried to make between stylistics and corpus linguistics on the one hand and sociolinguistics on the other. An urgent question is how to embed our findings in a sociolinguistic framework that can handle the systematic stylistic differences we discovered, especially between men and women.

References

Argamon, S., M. Koppel, J. Fine and A. Shimony, 2003. Gender, genre, and writing style in formal written texts. *Text*, 23 (3)

Arnaud, P. J. L., 1984. The lexical richness of L2 written productions and the validity of vocabulary tests. In T. Culhane, C. Klein-Bradley and D. Stevenson, eds., *Practice and Problems in Language Testing. Papers from the International Symposium on Language Testing*. Colchester: University of Essex, 14–28

Baayen, R. H., 1996. The effect of lexical specialization on the growth curve of the vocabulary. *Computational Linguistics*, 22: 455–480

Baayen, R. H., 2009. Corpus linguistics in morphology: morphological productivity. In A. Luedeling and M. Kyto, eds., *Corpus Linguistics. An international handbook*. Mouton De Gruyter, Berlin, 900–919

Biber, D., 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge

Biber, D., 1995. *Dimensions of Register Variation*. Cambridge University Press, Cambridge

Biber, D. and S. Conrad, 2009. *Register, Genre, and Style*. Cambridge University Press, Cambridge

Burrows, J. F., 1992a. Computers and the study of literature. In C. S. Butler, ed., *Computers and Written Texts*. Blackwell, Oxford, 167–204

Burrows, J. F., 1992b. Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7: 91–109

Burrows, J. F., 1993a. Noisy signals? Or signals in the noise? In *ACH-ALLC Conference Abstracts*. Georgetown, 21–23

Burrows, J. F., 1993b. Tiptoeing into the infinite: Testing for evidence of national differences in the language of English narrative. In S. Hockey and N. Ide, eds., *Research in Humanities Computing '92*. Oxford University Press, London

Eckert, P. and J. R. Rickford, eds., 2011. *Style and Sociolinguistic Variation*. Cambridge University Press, Cambridge

Geeraerts, D., S. Grondelaers and D. Speelman, 1999. *Convergentie en Divergentie in de Nederlandse Woordenschat. Een Onderzoek naar Kleding- en Voetbaltermen*. Meertens Instituut, Amsterdam

Grondelaers, S., K. Deygers, H. van Aken, V. van den Heede and D. Speelman, 2000. Het CONDIV-corpus geschreven Nederlands. *Nederlandse Taalkunde*, 5: 356–363

Grondelaers, S., H. van Aken, D. Speelman and D. Geeraerts, 2001. Inhoudswoorden en preposities als standaardiseringsindicatoren: De diachrone en synchrone status van het Belgisch Nederlands. *Nederlandse Taalkunde*, 6: 179–202

Halliday, M. A. K., 1964. Comparison and translation. In M. Halliday, M. McIntosh and P. Strevens, eds., *The linguistic sciences and language teaching*. Longman, London

Halliday, M. A. K., 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Edward Arnold, London

Halliday, M. A. K. and R. Hasan, 1976. *Cohesion in English*. Longman, London

Halliday, M. A. K. and C. M. I. M. Matthiessen, 2004. *An Introduction to Functional Grammar*. Longman, London, third, revised edition

Härnqvist, K., U. Christianson, D. Ridings and J.-G. Tingsell, 2003. Vocabulary in interviews as related to respondent characteristics. *Computers and the Humanities*, 37: 179–204

Heylighen, F. and J.-M. Dewaele, 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7 (3): 293–340

Jacobi, I., 2009. On variation and change in diphtongs and long level vowels of spoken Dutch. Ph.D. thesis, University of Amsterdam

Johansson, V., 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. In *WorkingPapers*, volume 53. Lund University, Dept. of Linguistics and Phonetics, 61–79

Joos, M., 1961. *The Five Clocks*. Harcourt, Brace and World, New York

Labov, W., 1966. *The social stratification of English in New York City*. Centre for Applied Linguistics, Washington, D.C.

Newman, M. L., C. J. Groom, L. D. Handelman and J. W. Pennebaker, 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45: 211–236

Oostdijk, N. H. J., 2002. The Design of the Spoken Dutch Corpus. In P. Peters, P. Collins and A. Smith, eds., *New Frontiers of Corpus Research*. Rodopi, Amsterdam, 105–112

Plevoets, K., 2008. Tussen spreek- en standaardtaal. Een corpusgebaseerd onderzoek naar de situationele, regionale en sociale verspreiding van enkele morfosyntactische verschijnselen uit het gesproken Belgisch-Nederlands. Ph.D. thesis, Katholieke Universiteit Leuven

Quené, H., 2008. Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *Journal of the Acoustical Society of America*, 123 (2): 1104–1113

Reid, T. B., 1956. *Linguistics, Structuralism, Philology*, *Archivum Linguisticum*, volume 8. Jackson, Son & Company, Glasgow

Richards, B., 1987. Type/Token ratios: What do they really tell us? *Journal of Child Language*, 14: 201–209

Schuppler, B., 2011. Automatic analysis of acoustic reduction in spontaneous speech. Ph.D. thesis, Radboud University Nijmegen

Tweedie, F. J. and R. H. Baayen, 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32: 323–352

Van Bergen, G., W. Stoop, J. Vogels and H. de Hoop, 2011. Leve hun! Waarom hun nog steeds hun zeggen. *Nederlandse taalkunde*, 16 (1): 2 – 29

Van de Ven, M., M. Ernestus and R. Schreuder, submitted. Predicting words in spontaneous speech: The role of context

Van Dijk, T. A., 2009. *Society and Discourse: How social contexts influence text and talk*. Cambridge University Press, Cambridge

Van Gijsel, S., 2007. Sociovariation in Lexical Richness. A Quantitative Corpus Linguistic Analysis. Ph.D. thesis, Katholieke Universiteit Leuven

Van Marle, J., 1988. Betekenis als factor bij produktiviteitsverandering. *Spektator*, 17: 341–359

CHAPTER 2

Variation in Dutch: From written MOGELIJK to spoken MOK[1]

Abstract

In Dutch, high-frequency words with the suffix *-lijk* are often highly reduced in spontaneous unscripted speech. This study addressed socio-geographic variation in the reduction of such words against the backdrop of the variation in their use in written and spoken Dutch. Multivariate analyses of the frequencies with which the words were used in a factorially contrasted set of subcorpora revealed significant variation involving the speaker's country, sex, and education level for spoken Dutch, and involving country and register for written Dutch. Acoustic analyses revealed that Dutch men reduced most often, while Flemish highly educated women reduced least. Two linguistic context effects emerged, one prosodic, and the other pertaining to the flow of information. Words in sentence final position showed less reduction, while words that were better predictable from the preceding word in the sentence (based on mutual information) tended to be reduced more often. The increased probability of reduction for forms that are more predictable in context, combined with the loss of the suffix in the more extremely reduced forms, suggests that high-frequency words in *-lijk* are undergoing a process of erosion that causes them to gravitate towards monomorphemic function words.

[1]This study, co-authored by Mirjam Ernestus, Roeland van Hout and Harald Baayen, is published under the same title in *Corpus Linguistics and Linguistic Theory* 1 – 2 (2005), 183 – 223.

2.1 Introduction

In spontaneous speech words are often pronounced in reduced form (Ernestus, 2000; Johnson, 2004). Some words are reduced to such an extent that an faithful orthographic transcription would be very different from the orthographic norm. An example from Dutch is the word *mogelijk* ('possible'), which can be pronounced not only as [moxələk] but also as [moxək], [molək], or even as [mok].

Strongly reduced word forms are difficult to interpret without syntactic or semantic context (Ernestus et al., 2002). When speakers of Dutch are presented with the word [mok] in isolation, they find it difficult to assign a meaning to this string of phonemes. It is only when the word is embedded in a sentence that its meaning becomes available. Interestingly, listeners who understood the meaning of [mok] tend to think they heard the full, unreduced form [moxələk] (Kemps et al., 2004). A central question in the research on the comprehension of reduced words is what aspects of the linguistic context allow the listener to access the associated semantics.

An important predictor for the degree of reduction in speech production is lexical frequency, as demonstrated by Jurafsky et al. (2001) for function words. The more often a function word is used in speech, the more likely it is to undergo reduction, in line with Zipf's law of abbreviation (Zipf, 1935). Bybee (2001) discussed how frequency of occurrence affects the realization of word final dental plosives in monomorphemic words. Pluymakers et al. (2005) observed a negative correlation between frequency and acoustic length for several kinds of derived words in Dutch, including words with the suffix *-lijk*, the suffix in the above example *moge-lijk*. Jurafsky et al. also showed that the degree of reduction is modulated by the extent to which a word is predictable from its context. However, it is currently an open question to what extent the use of reduced forms is codetermined by socio-geographic factors.

Various corpus-based studies have shed light on variation in language use in general. Biber (1988, 1995) identified different varieties of English (and also other languages) by means of factor analyses of the frequencies of a broad range of morphological and syntactic variables. In the domain of literary studies, Burrows (1992a, 1986, 1987, 1992b, 1993a,b) demonstrated regional variation in English narrative, diachronic change in literary texts, and even sex-specific differences in the writing of English historians born before 1850 on the basis of the most common words. Studies in authorship attribution revealed, furthermore, that differences in speech habits can sometimes be traced down to the level of individual language users (Holmes, 1994; Baayen et al., 1996, 2002). Finally, Baayen (1994) and Plag et al. (1999) showed that derivational affixes are used to a different extent in spoken and written registers.

The aim of the present study is to investigate the extent to which the use of words in *-lijk* varies systematically in both written and spoken Dutch. Words in *-lijk* are generally classified as open-class words. However, it is noteworthy that the suffix *-lijk* is hardly productive (Van Marle, 1988), and that many

high-frequency forms are no longer semantically compositional. For instance, *natuur-lijk*, literally 'nature-like', usually means 'of course'. In this study, we will first investigate systematic variation of this unproductive suffix in written Dutch as function of whether a text is written in Flanders or in the Netherlands, and as a function of its register. Second, we explore spoken Dutch as a function of whether a speaker lives in Flanders or in the Netherlands, of the speaker's sex, and the speaker's level of education. Third, we address the question to what extent reduction in the acoustic form of words in *-lijk* is predictable from socio-geographic variables.

In this study, we have made extensive use of multilevel analysis of covariance, a statistical technique that offers two advantages compared to principal components analysis, factor analysis, and correspondence analysis (Lebart et al., 1998). First of all, multilevel modeling allows the researcher to directly assess the significance of predictors, as well as how individual words (or other units of analysis) interact with these predictors. In other words, instead of using both a clustering technique such as principal components analysis and a technique for group separation such as discriminant analysis, we were able to fit a single statistical model to the data that allows us both to trace what predictors are significant, and to visualize their effects. The second advantage of multilevel modeling is that it offers the researcher the possibility to include covariates (such as mutual information) in the model.

2.2 Written Dutch

For our study of written Dutch, we made use of the CONDIV corpus (Grondelaers et al., 2000). This corpus comprises three kinds of written Dutch: written Dutch from newspapers, written Dutch from USENET, and written Dutch from chat sites. In the present study, we investigated lexical variation in the subcorpus of newspapers. The CONDIV corpus sampled four Flemish newspapers (*De Standaard*, *Het Laatste Nieuws*, *De Gazet van Antwerpen* and *Het Belang van Limburg*) and three Dutch newspapers (*NRC Handelsblad*, *De Telegraaf* and *De Limburger*). These seven newspapers can also be cross-classified according to their register. *De Standaard* and *NRC Handelsblad* are Quality newspapers, aiming at a more educated readership. *Het Laatste Nieuws* and *De Telegraaf* are National newspapers, and *De Gazet van Antwerpen*, *Het Belang van Limburg*, and *De Limburger* are Regional newspapers.

For each of the seven newspapers in the CONDIV corpus, we selected the first 1.5 million words (the size of the smallest newspaper) for further analysis. From these data sets, we selected the 80 most frequent words in *-lijk* (listed in the appendix) that occurred at least once in each of the seven subcorpora, and registered their frequencies in these subcorpora, which we cross-classified by Country and Register. (Pooling the most common words in each of the subcorpora separately led to a change in only one word.) In this way, we obtained a table with 80 rows (words) and 7 columns (newspapers). One way of look-

ing at these data is that the seven newspapers are represented as 7 points in an 80-dimensional space. This raises the question whether the way in which these seven newspapers are distributed in this space reflects the Registers and Countries of these newspapers.

There are many different statistical techniques for addressing this question, among which principal component analysis, factor analysis, and correspondence analysis are currently the most widely used. Each of these techniques allows the researcher to explore the structure among our newspapers by means of dimension reduction. Figure 2.1 summarizes the results of a principal component analysis. The left panel plots the newspapers in lexical space by means of the first two principal components. The first principal component (PC1) accounted for 37.3% of the variance, the second (PC2) accounted for 20.1% of the variance. As can be seen in the left panel of Figure 2.1, these two components reflect the geographic and register differences between the newspapers. First consider PC1. The Flemish newspapers, represented in upper case letters, occur more to the left of the graph, while the Dutch newspapers appear more to the right. In other words, PC1 captures the geographical variation in the use of the 80 high-frequency words in *-lijk* that we sampled. PC2, on the other hand, captures aspects of the register variation. The Quality newspapers (*NRC Handelsblad*, denoted by `nrc` in the plot, and *De Standaard*) appear lower in the plot, while the National newspapers, *Het Laatste Nieuws* and *De Telegraaf* appear at the top of the graph. In the right panel the loadings of the target words on the newspapers is plotted. Words positioned lower in the plot, for instance, have the highest load on *De Standaard*, and are thus most often used in that newspaper.

In order to ascertain to what extent this interpretation is statistically robust, we carried out two tests contrasting the coordinates of the newspapers on the two principal components. A Welch Two Sample t-test contrasting the Flemish and Dutch newspapers with respect to PC1 revealed a highly significant difference ($t(4.16) = -8.47$, $p = 0.0009$), and a one-way analysis of variance contrasting the three Registers with respect to PC2 also revealed significant differences ($F(2, 4) = 8.07$, $p = 0.0394$).

Although these tests support the conclusions we drew from the visual inspection of Figure 2.1, there are a number of questions that this exploratory technique does not answer. One of these questions concerns the possibility of an interaction between Country and Register. Do these two factors work independently, or might the effect of one of these factors be different depending on the value of the other factor? Second, are these geographic and register differences supported in the same way by each of our 80 words? It might be the case that the main effects uncovered by the principal components analysis are supported only by specific subsets of words. More technically, we would like to be able to ascertain whether there are interactions between the words and Register and Country. We therefore analyzed these data in more detail using multilevel regression modeling.

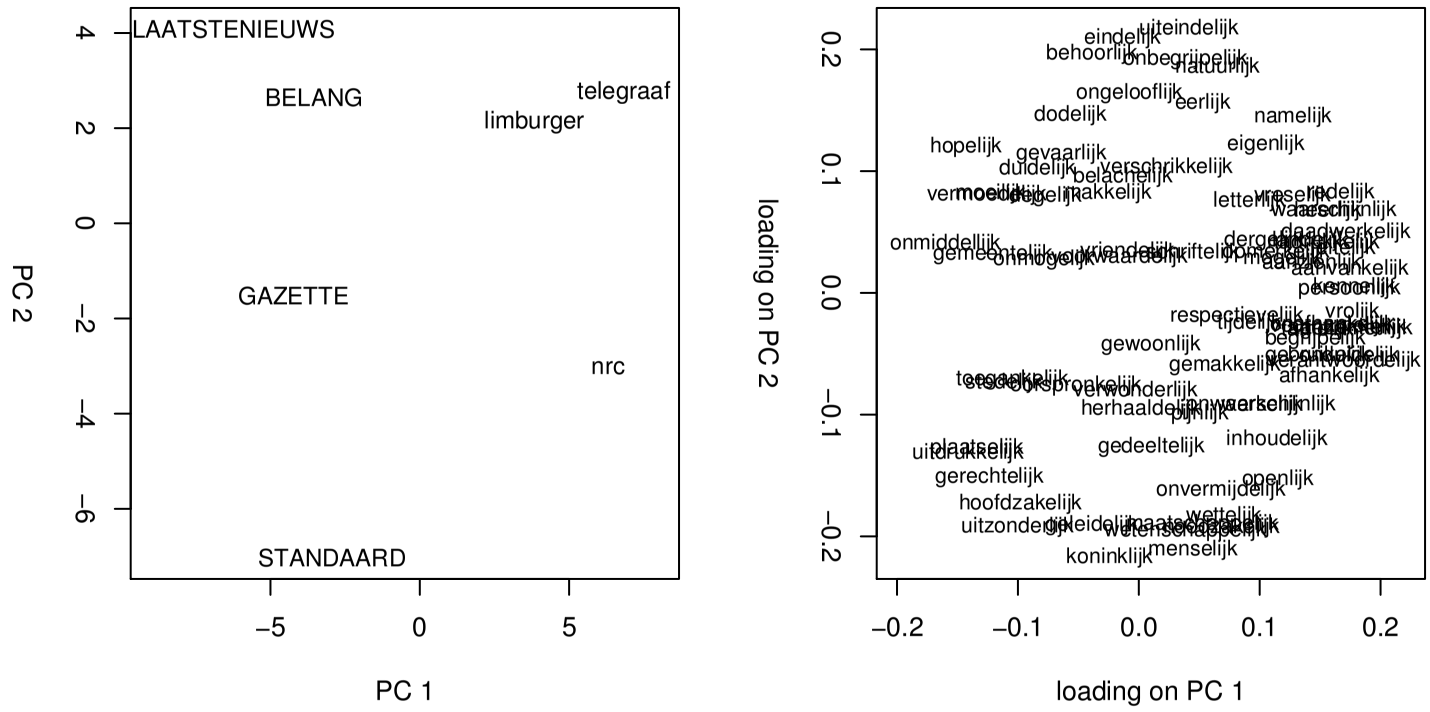


Figure 2.1: Principal Component Analysis of 80 words ending in *-lijk* in the seven (CONDIV) newspapers. The names of Flemish newspapers are in capital letters.

Multilevel modeling (Pinheiro and Bates, 2000) is a regression technique developed to deal specifically with data combining fixed and random effects. Factors are described as 'fixed' when the levels of that factor exhaust all possible levels. An example of a fixed effect in the present data is Country: the Netherlands and Flanders are the only two European countries in which Dutch is spoken, there are no other conceivable levels of this factor that we have not sampled. By contrast, the words in our data set constitute a 'random' effect: these words are sampled from a larger population of words in *-lijk*, and we would like to know whether the patterns observed in the data would generalize to the whole class of words in *-lijk*. In the model that we fit to these data, we therefore included Word as a random factor, it is the main grouping factor in the analyses to follow. Mixed effects models deal with the distinction between fixed and random effects in a more principled way than do traditional linear models, and, more importantly, they provide more precise estimates of the random effects (in this study, improved estimates of the effects of the individual words). In addition, these by-word adjustments are easier to extract and inspect than with standard or general linear models (Quené & Van den Bergh, 2004; Baayen, 2004).

Recall that we have 7 observations for each word, one frequency count for each newspaper. One way of looking at what multilevel modeling does is to build informed models for each of the individual words. The individual models are informed in the sense that they are constructed against the background of what is known about the behavior of all the other words in the sample.

A multilevel model fit to the logarithmically transformed frequencies of the 80 words in *-lijk* in the seven newspapers (using a stepwise model selection procedure), with Word as grouping factor, revealed a significant (fixed) effect for Country ($F(1, 463) = 9.3067$, $p = 0.0024$), a marginally significant (fixed) effect for Register ($F(2, 463) = 2.4592$, $p = 0.0866$), and a significant interaction of Country by Register ($F(2, 463) = 16.1930$, $p < 0.0001$). The frequencies of words in *-lijk* tended to be lower in Flanders compared to the Netherlands. In both countries, words in *-lijk* were used most frequently in the Quality newspaper. Furthermore, in Flanders words in *-lijk* were used significantly less often in the National newspaper than in the Quality newspaper. Conversely, in the Netherlands words in *-lijk* were used significantly less often in the Regional newspaper than in the Quality newspaper. This model provides further support for the general patterns discovered by the principal components analysis. However, it also provides a correction by uncovering an interaction of Country by Register. In addition, the multilevel model points not only to a difference between Flanders and the Netherlands with respect to the use of words in *-lijk*, but also discloses that, apparently, words in *-lijk* are used slightly more often in the Netherlands.

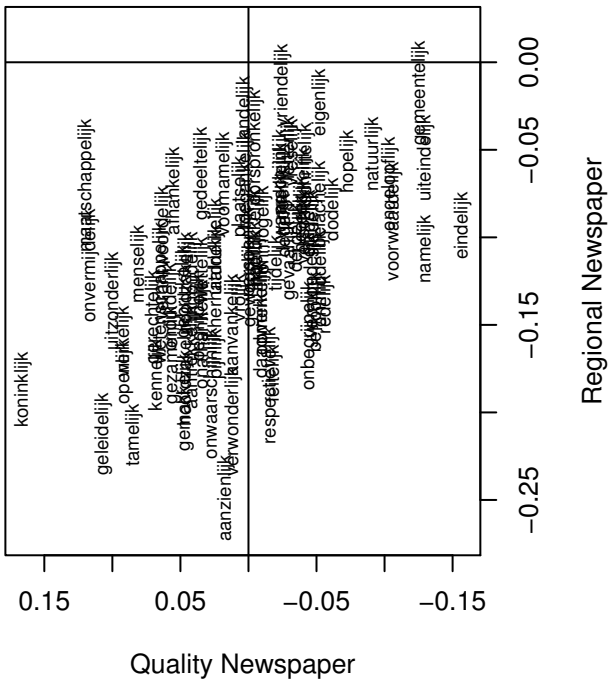
In multilevel modeling, it is also possible to investigate whether there are interactions between the fixed effects and the main grouping factor Word. We observed significant interactions involving Word both for Country and for Reg-

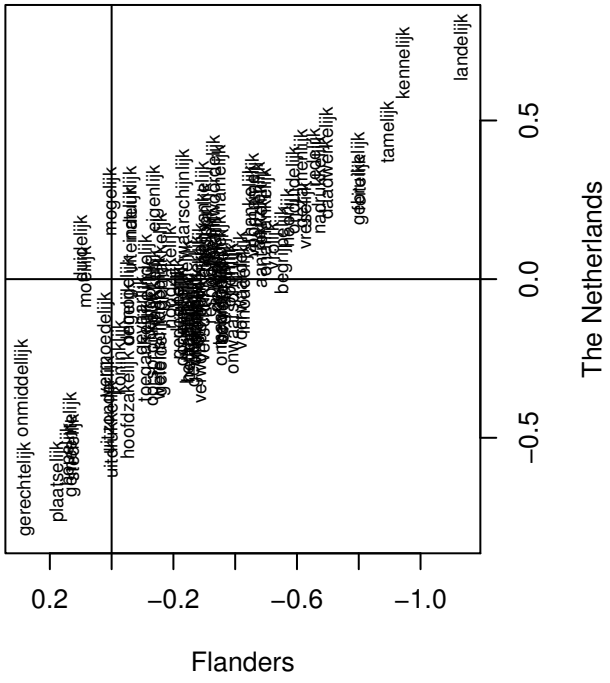
ister ($p < 0.0001$ and $p < 0.0023$, likelihood ratio tests). There are two further technical details concerning this model. First, we removed outliers from the data set, i.e., data points with standardized residuals with an absolute value exceeding 2 standard deviation units (see Chatterjee et al., 2000 for further details on the removal of outliers in multiple regression). In the present model this led to the removal of 12 data points (2.1% of the 560 data points). Second, we added an extra parameter to the model in order to remove the heteroscedasticity visible in the plot of the standardized residuals against the fitted values. This extra parameter (for an exponential variance function, see Pinheiro & Bates 2000, 211-213) was also justified by a likelihood ratio test ($p < 0.0001$).

Figure 2.2 provides a visual aid to understanding the interactions involving Word. The upper left panel shows the interaction of Word by Country. Recall that we observed a main effect for Country, with words in *-lijk* being used more frequently in the Netherlands. The interaction of Country by Word shows that this effect is not equally strong for all words. The horizontal axis of the upper left panel shows the by-word adjustments that need to be made in order to make the predictions for the frequencies of the words as used in the Netherlands more precise. The vertical axis does the same for the predictions pertaining to the Flemish frequencies. Positive values indicate that a word is used more often than the average word in *-lijk* in the country associated with the axis. In other words, the further to the right a word is positioned, the more frequently it is used in the Netherlands. The higher a target is positioned, the more frequently it is used in Flanders. The words, *landelijk* ('national') and *kennelijk* ('apparently'), for instance, are used more often in the Netherlands than in Flanders, while *onmiddellijk* ('immediately'), and *gerechtelijk* ('judicial') are used more often in Flanders.[1] We listed the coordinates of all words in Figure 2.2 as well as the coordinates of all words in the following figures in Appendix B.

A closer inspection of this plot and the corresponding table of by-word adjustments suggests that the locatives *gemeentelijk* ('municipal'), *plaatselijk* ('local') and *stedelijk* ('urban') are used more frequently in Flanders while the locative *landelijk* ('national') is used more frequently in the Netherlands. Moreover, there are two near-synonyms for *explicit(ly)* that show differential use across the two countries: *Uitdrukkelijk* is typically Flemish and *nadrukkelijk* is typically Dutch.

[1] Multilevel models only specify whether an interaction involving the main grouping factor (Word in the present example) is significant, but do not provide means for comparing the significance of differences involving individual words. Questions such as whether a given word occurs significantly more often in Flanders or in the Netherlands require independent statistical tests, for instance, tests based on contingency tables such as Fisher's exact test of independence. Note that such independent tests are justified only in the present framework for comparisons for which significant interactions with the main grouping factor have been observed.


koninklijk
geleidelijk
tamelijk
menselijk
onvermijdelijk
eigenlijk
hopelijk
natuurlijk
namelijk
eindelijk
0.15
0.05
−0.05
−0.15
−0.25
0.00
−0.05
−0.15
Quality Newspaper
Regional Newspaper


gerechtelijk onmiddellijk
plaatselijk
modelijk
mogelijk
tamelijk
kennelijk
landelijk
0.2
−0.2
−0.6
−1.0
0.5
0.0
−0.5
Flanders
The Netherlands

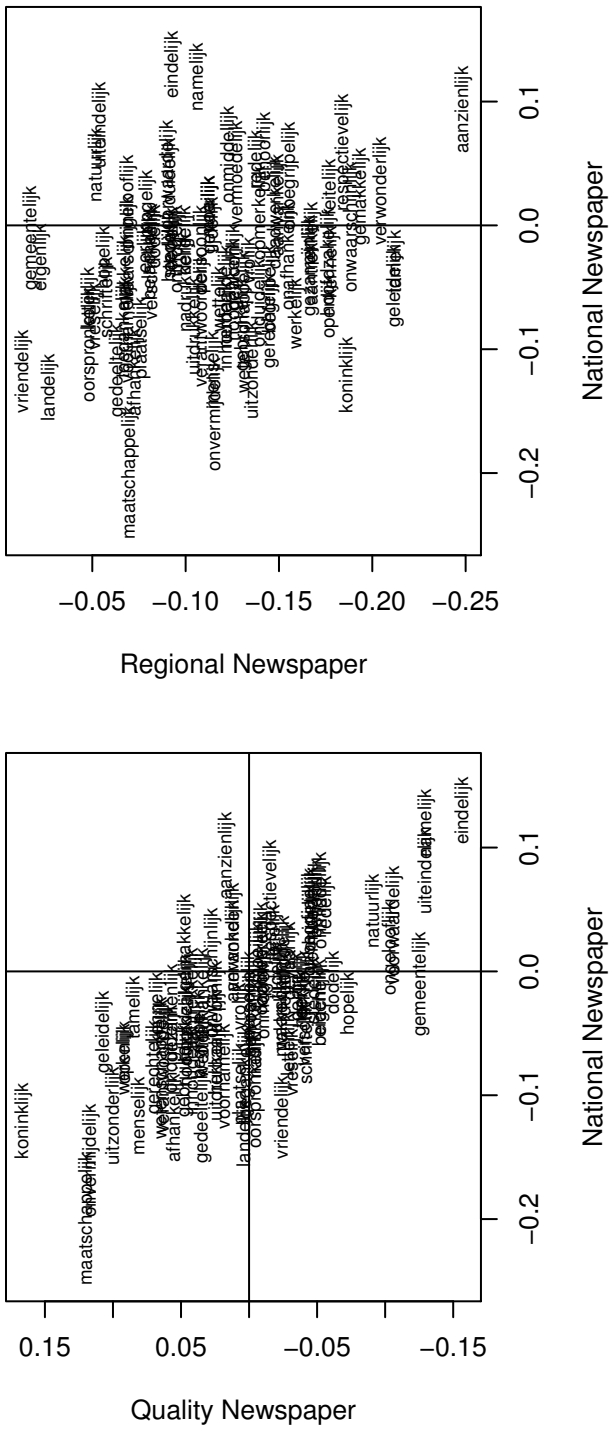


Figure 2.2: By Word adjustments for Country and Register in a multilevel model for 80 selected words ending in *-lijk* from the seven CONDIV newspapers.

The remaining three panels of Figure 2.2 plot the register variation for words in *-lijk*. The upper right panel, for instance, shows the variation of use of words in *-lijk* for Regional compared to Quality newspapers. For example, the word *gemeentelijk* ('municipal') appears more frequently in Regional newspapers, and less frequently in Quality newspapers. Words typical for the Quality newspapers are, among others, *koninklijk* ('royal'), *onvermijdelijk* ('inevitable'), and *geleidelijk* ('gradual'). A word typical for the Regional newspapers is *gemeentelijk* ('municipal'). Note that most words appear more frequently in the Quality newspapers than in the Regional newspapers.

The question that arises at this point is whether the geographic and register variation in the use of words in *-lijk* is specific to these particular complex words, or whether this variation is also reflected in the use of other aspects of lexis and grammar. In other words, we need an independent and established method for tracing variation in other parts of grammar and lexis in order have a benchmark with which the present results can be compared.

The benchmark that we selected is the stylometric technique developed by Burrows (1992a, 1993a). Burrows showed that differences in speech habits of individual language users are reflected in their use of the most common word types. The most common words typically include function words (determiners, pronouns, conjunctions, auxiliaries) as well as some common adverbs. Differences in the use of the most common words tend to represent differences in syntactic habits (Baayen et al., 1996). Content words are usually excluded from the list of most common words in stylometric studies, in order to avoid clustering based on topical rather than on structural linguistic features. We applied this state-of-the-art approach from stylometry not at the level of individual speakers but at the aggregate level of groups of speakers defined by socio-geographic variables. We used the same corpus of Dutch and Flemish newspapers, and selected the 80 most common words, excluding 3 content words from this list. These words are listed in the appendix.

A multilevel model fit to the logarithmically transformed frequencies of the 80 most common words that appeared in each of the seven newspapers revealed significant main effects for Country ($F(1, 463) = 41.478$, $p < 0.0001$), Register ($F(2, 463) = 50.854$, $p < 0.0001$) and an interaction of Register by Country ($F(2, 463) = 45.168$, $p < 0.0001$). There were no significant differences pertaining to Register within the set of Dutch newspapers. Within the set of Flemish newspapers, the Regional newspapers used the 80 most common words equally often as the Dutch newspapers, in contrast to the Quality newspapers, which used them least often. Similar to the case of the words in *-lijk*, we observed significant interactions between the main grouping factor (Word) and Country as well as Register (both $p < 0.0001$, likelihood ratio tests). This model was obtained after removing 12 influential outliers (2.1% of the 560 data points). We again used an exponential variance function in order to remove heteroscedasticity visible in the plot of the standardized residuals against the fitted values. As before, this involved adding an additional parameter to the model, which

was justified by a likelihood ratio test ($p < 0.0001$).

When we collapse over the different registers, we find that the most common words in the present study are used less often in Flanders than in the Netherlands. This is probably due to the selection of only the 80 most frequent common words for analysis. The dialects of Flanders are characterized by a much greater variety of forms than those in the Netherlands, especially in the pronominal system. Furthermore, the standard language in Flanders is more divorced from the language varieties used in informal communicative situations compared to the Netherlands. We suspect that some dialectal variants were used in the Flemish materials along with the standard forms. If so, the standard forms were used somewhat less frequently than in the corresponding Dutch texts.

Interestingly, the interaction of Country by Register shows that this difference between Dutch and Flemish emerges most markedly for the Quality Belgian newspaper *De Standaard*. Since this journal is known to use a rather formal style, the low frequency of most common words in this journal cannot be ascribed to the presence of dialectal forms. It is more likely that this difference suggests that the journalists writing for this newspaper use more content words in their articles than journalists writing for the other newspapers, which leads to a higher information density.

Figure 2.3 illustrates the interaction of Word by Country. The x-axis shows the by Word adjustments of the relative frequency necessary to obtain an accurate estimate of the relative frequency of each word as used in the Netherlands. The y-axis shows the extra by Word adjustments needed to obtain the relative frequency of the words as used in Flanders. The word *ik* ('I'), for instance, is used more frequently in Flanders than the average most common word in the data set, as shown by its high value on the y-axis.

The way in which the different most common words are positioned suggests that the first person pronouns (*we* 'we', *ik* 'I') are used more often in Flanders, while third person pronouns are used more frequently in the Netherlands (*hij* 'he', *zij* and *ze* 'she', *zich* 'oneself'). Note that only three words are used more frequently in Flanders than in the Netherlands.

When we compare the socio-geographic variation observed for words in *-lijk* with the variation as indicated by the most common words, we find both similarities and differences. Both sets of words emerged as carriers of socio-geographic differentiation. Furthermore, both the most common words and the words in *-lijk* were used somewhat less often in Flanders. With respect to register, however, the two sets led to different results: the most common words were atypical for quality papers in Flanders, while the words in *-lijk* were more characteristic of quality newspapers in both regions.

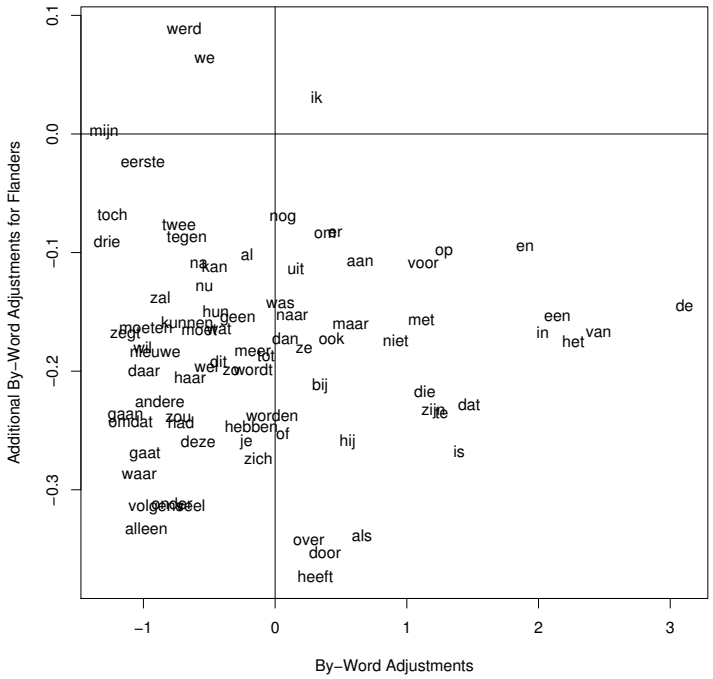


Figure 2.3: By Word adjustments for Country in a multilevel model for 80 selected most common word types from the seven CONDIV newspapers.

2.3 Spoken Dutch

Next we explored effects of socio-geographic variation on the frequency with which words in *-lijk* and most common words are used in spoken Dutch.

We made use of the Corpus of Spoken Dutch (CGN) (Oostdijk, 2002). This corpus contains approximately 8.9 million words of spoken Dutch, sampled from a wide range of registers. In order to maximize the contrast between written and spoken Dutch, we focused on the subcorpora containing recordings of spontaneous speech. The CGN comprises two categories of spontaneous Dutch: face-to-face conversations and telephone dialogues, in all 4,7 million words. The CGN provides detailed information about the different speakers including the country in which they live, their education level, and their sex. This made it possible for us to not only investigate the effects of Country (the Netherlands versus Flanders), but also the effects of Education (high (attended bachelor or master education) versus non high education level), and Sex (men versus

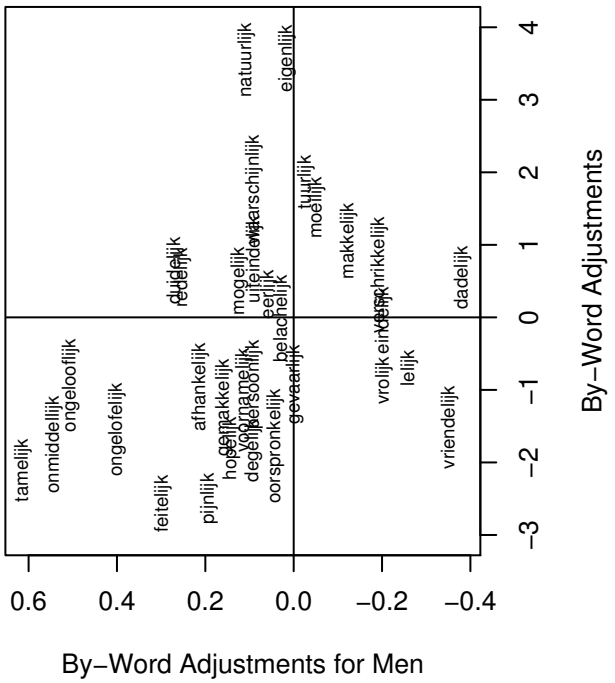
women).

We created eight subcorpora according to a $2 \times 2 \times 2$ factorial design with as factors Country, Sex, and Education. These subcorpora differed substantially in size, ranging from 189,000 words (for Flemish male speakers with a non high education level) to 1,200,000 words (for Dutch female speakers with a high education level). We then selected all words in *-lijk* that appeared at least once in each of these eight subcorpora (32 words, see appendix), and we calculated their relative frequencies in each subcorpus. These relative frequencies were the dependent variable in a multilevel model with Word as main grouping factor and Country, Sex, and Education as predictors.

In contrast to the results of our study of words in *-lijk* in written Dutch, the model fit to the logarithmically transformed relative frequencies revealed no significant main effect for Country. There was also no significant main effect for Sex. However, speakers with a higher education level tended to use words in *-lijk* more often than speakers with lower education levels ($F(1, 218) = 4.0514$, $p = 0.0454$). The main effect of Education in spoken Dutch mirrors the greater use of *-lijk* in the Quality newspapers as compared to the National newspapers, with as main difference that in spoken Dutch this simple main effect is not modulated further by an interaction with Country.

Furthermore, we observed significant interactions of Word by Country, Word by Sex, and Word by Education ($p < 0.001$, likelihood ratio tests). This model was obtained after removing six influential outliers (2.3% of the 256 data points). The heteroscedasticity visible in the plot of the standardized residuals against the fitted values was again brought under control with an additional parameter for the variance function, justified by a likelihood ratio test ($p < 0.0001$).

Figure 2.4 illustrates these interactions between Word and Country, Sex, and Education. All x-axes show the by-words adjustments necessary to obtain the relative word frequencies for Dutch women with a high education level. The y-axis of the upper left panel shows the extra adjustment of the relative frequency required for each word to obtain an accurate estimate for the relative frequency in Flanders (Flemish highly educated women). The words *eigenlijk* ('actually') and *natuurlijk* ('of course'), for instance, are used very frequently in the Netherlands. However, *eigenlijk* is used even more frequently in Flanders, while *natuurlijk* is used somewhat less frequently in Flanders.


tamelijk
onmiddellijk
ongelooflijk
ongelofelijk
feitelijk
pijnlijk
afhankelijk
oorspronkelijk
redelijk
mogelijk
natuurlijk
eigenlijk
moeilijk
makkelijk
vrolijk
lelijk
vriendelijk
dadelijk
By−Word Adjustments
By−Word Adjustments for Men

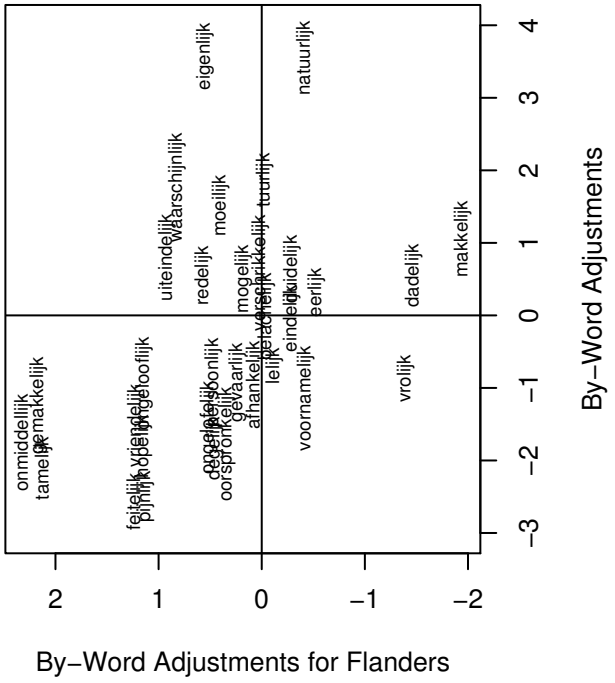

onmiddellijk
tamelijk
makkelijk
uiteindelijk
redelijk
moeilijk
eigenlijk
natuurlijk
vrolijk
dadelijk
makkelijk
By−Word Adjustments
By−Word Adjustments for Flanders

Figure 2.4: By Word adjustments for Country, Sex, and Education in a multilevel model for 32 selected words ending in *-lijk* from the eight, factorially designed, subcorpora of the spontaneous conversations in the CGN.

In addition, this panel shows that *onmiddellijk, gemakkelijk* and *tamelijk* ('immediately', 'easily', 'somewhat') are typical for Flanders, while *vrolijk, dadelijk* and *makkelijk* ('happy', 'immediately', 'easily') are typical for the Netherlands. Interestingly, *onmiddellijk* and *dadelijk* are (near) synonyms for 'immediately', and *gemakkelijk* and *makkelijk* are variants of 'easily'. It is standard practice in sociolinguistics to investigate linguistic variation in time and space by means of pairs of expressions that differ in one dimension only. At the lexical level, this implies that only pairs such as *gemakkelijk/makkelijk* and *dadelijk/onmiddellijk* would be used to probe sociolinguistic variation at the lexical level. What the methodology explored in the present study allows us to observe is that such matched pairs of words indeed are strong carriers of variation, but that there are other, non-matched words such as *tamelijk* ('somewhat') and *vrolijk* ('happy') that are also involved in this geographical opposition.

The upper right panel shows the by-word effects for Sex, with on the y-axis the adjustment for frequency of use for men (Dutch highly educated men). The near synonyms of 'immediately', *onmiddellijk* (men) and *dadelijk* (women) are clear markers for the two sexes, but, as before, there are also other, non-synonymous markers, such as *tamelijk* and *ongelooflijk* ('somewhat', 'unbelievable', more typical for men) versus *vriendelijk* and *lelijk* ('friendly' and 'ugly', more typical for women).

The lower left panel illustrates the required adjustment of the relative frequency of the different words for non highly educated speakers (Dutch non-highly educated women). It shows that the synonyms of 'immediately' also differentiate between the education level of speakers, together with *vriendelijk, lelijk, vrolijk* ('friendly', 'ugly', 'happy' for non high education) and *tamelijk, ongelooflijk* ('somewhat', 'unbelievable' for high education).

As in the analyses of written Dutch, we investigated whether the socio-geographic variation in the use of *-lijk* is also reflected in the use of the most common words. A multilevel model fit to relative frequencies, raised to the power of 0.25[2], of the 80 most common words occurring in all eight subcorpora (see appendix) revealed only one significant main effect: Men tended to use the 80 most common words less often than women ($F(1, 551) = 14.759$, $p = 0.0001$). This suggests that the speech of men is characterized by a slightly higher information density compared to women. (This higher information density may be due to more intensive use of less common non-content words, but also to the use of more content words.) We observed significant interactions between the main grouping factor Word, and Country, Sex, and Education (all $p < 0.0001$, likelihood ratio tests). This model was obtained after the removal of eight influential outliers (1.3% of the 640 data points).

The interaction of Word by Country is illustrated in Figure 2.5. Again, the x-axis shows the by Word adjustments to obtain the relative frequency of the different words as used in the Netherlands, and the y-axis shows the extra ad-

[2]This transformation brought the distribution of relative frequencies more in line with the normality assumptions underlying linear regression.

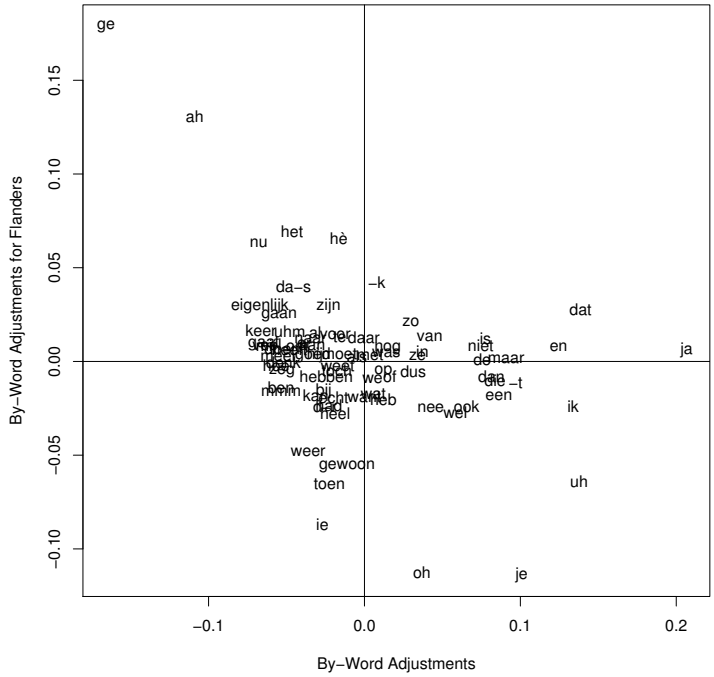


Figure 2.5: By Word adjustments for Country in a multilevel model for the 80 most common words from the eight, factorially designed, subcorpora of the spontaneous conversations in the CGN.

justment of the relative frequency for each word to obtain the relative frequency of the word as used in Flanders. The near-synonyms *ge* ('you', Flanders) and *je* ('you', the Netherlands) are clear markers of the differences in lexis across countries. Furthermore, the discourse marker *ah* appears to be preferred by Flemish speakers, and *oh* by Dutch speakers.

Summing up, in both spoken and written Dutch, and for both words in *-lijk* and most common words, all predictors interacted with Word. Furthermore, there were differences in the main effects. For words in *-lijk*, we observed that speakers with a higher education level used these words more often, and so did the Dutch Quality newspaper. The selected most common words in spoken Dutch were used less frequently by men than by women. Furthermore, the most common words selected from written texts were used more frequently in the Netherlands compared to Flanders, and in Flanders differentiated between the different kinds of newspapers.

2.4 Variation in the reduction of *-lijk*

The preceding analyses of *-lijk* in spoken Dutch proceeded on the basis of the orthographic transcriptions of spontaneous conversations. These analyses glossed over a property of these words that is a potential carrier of socio-geographic differences, namely, the extent to which these words are reduced acoustically in casual speech.

In order to explore this potential socio-geographic stratification of acoustic reduction, we selected those words in *-lijk* that occurred more than 75 times in the subcorpus of spontaneous Dutch from the set of 32 words in *-lijk* examined above. For these 24 words, we aimed at randomly selecting the acoustic signal for ten occurrences in each of the eight cells of the design obtained by factorially contrasting Country, Sex, and Education. In roughly one third of the cases it turned out to be impossible to obtain even ten occurrences, either because of data sparseness or because of a variety of problems with the acoustic signal itself. Instead of the desired 80 tokens for each of the 24 selected words the mean number of observations for a word in our design was 64.3, the median was 64 and the range was 43 to 80. The total number of observations was 1543.

A broad phonological transcription, made by one transcriber, for each of these 1543 sound files served as the basis for assignment to one of three levels of Reduction: No Reduction, Medium Reduction, and High Reduction. Words were classified as having No Reduction either when both the suffix and the stem were fully preserved [moxələk] (mogelijk, 'possible'), [tyrlək] (tuurlijk, 'of course') or when the suffix *-elijk* was reduced to *lijk* and the stem was fully preserved [moxlək] (*mogelijk*). Reduction of the suffix from *-elijk* to *-lijk* was not classified as reduction for two reasons. The first reason was that both *-elijk* and *-lijk* are allomorphs of the same suffix. Which allomorph is used, depends only on the phonemes preceding the suffix. The second reason was that it was often hard to ascertain whether or not the schwa was still present in the suffix. Words were classified as having Medium Reduction when the /l/ from the suffix or when consonants from the coda of the stem were not present [molək], (*mogelijk*), [moxək], (*mogelijk*), [ɛmlək] (*eindelijk*), [ɛɪdlək] (*eindelijk, 'finally'*). If the coda of the stem had more than one consonant, one of these consonants and the /l/ of the suffix could be absent: [ɛɪmək], [ɛɪdək], [ɛɪlək] (all forms of *eindelijk*). Words were classified as having High Reduction either when the suffix was completely integrated with the stem, with the final /k/ of the suffix becoming the coda of the stem [mok] (*mogelijk*), [moxk] (*mogelijk*), or when the suffix had disappeared completely [mo] (*mogelijk*).

Of the 24 initially selected words, represented by 1543 tokens, only 14 words appeared in a Medium or High Reduced form: *afhankelijk* ('dependent'), *dadelijk* ('immediately'), *duidelijk* ('clear'), *eerlijk* ('honest/fair'), *eigenlijk* ('actually'), *eindelijk* ('finally'), *moeilijk* ('difficult'), *mogelijk* ('possible'), *natuurlijk* ('of course'), *persoonlijk* ('personal'), *tuurlijk* ('of course'), *uiteindelijk* ('finally'), *vriendelijk* ('friendly'), and *waarschijnlijk* ('probably'), in all 946

word tokens. In order to provide some validation for the three categories of reduction and the initial assignment of the word tokens to these categories, a second judge also listened to each of these 946 words and assigned them to one of the three reduction categories. For 19 tokens the new assignment deviated from the original one. A third judge determined the final assignment for these word tokens. We calculated two statistics for each of the words: the relative frequency of the word in a given subcorpus, and the mutual information (Church and Hanks, 1990; Gregory et al., 1999) of the word and the word preceding it, which estimates the predictability of a word given the preceding word in the sentence. (For words with a frequency less than 11 in a subcorpus, the mutual information was set to zero in order to avoid excessively high and uninformative mutual information values.) Finally, we registered whether a token occurred in the final or in a non-final position in the sentence.

In the preceding statistical analyses we used multilevel models with Word as main grouping factor. By modeling Word as a random effect, the results of the statistical analyses generalized to the population of (higher frequency) words from which we sampled our materials. Since we have only 14 words ending in *-lijk* in the present data set, and since these 14 words are in no way a random sample, we opted for analyzing Word as a fixed effect in the analyses to follow.

Six of the 14 words were characterized by only two levels of reduction (High Reduction versus No Reduction). For eight words, all three levels of reduction were attested in colloquial Dutch. In what follows, we analyzed the log odds ratio of the number of words with No Reduction to the number of words with High or Medium reduction, using logistic regression (Harrell, 2001).

A logistic simple main-effects model of covariance fitted to the 946 data points (using a stepwise model selection procedure) revealed significant effects for Country ($X^2_{(1)} = 13.15$, $p = 0.0003$), Sex ($X^2_{(1)} = 7.35$, $p = 0.0067$), Position ($X^2_{(1)} = 6.69$, $p = 0.0097$), Mutual Information ($X^2_{(1)} = 7.83$, $p = 0.0051$), and Word ($X^2_{(13)} = 235.10$, $p < 0.0001$). When we allowed two-way interactions into the model, interactions emerged of Country by Education ($X^2_{(1)} = 6.09$, $p = 0.0136$) and of Country by Word ($X^2_{(13)} = 26.12$, $p = 0.0164$). Due to the latter interaction, the main effect of Country, which revealed that speakers in Flanders reduced less than speakers in the Netherlands, was no longer significant. Thus, it appeared that words in *lijk* are overall more often reduced by Dutch speakers, but that this is not the case for all the individual words. The coefficients of this model, with the exception of those involving the interaction of Country by Word, are summarized in Table 2.1. The generalized R^2 index for this model was 0.568, and Somer's D_{xy} was 0.778.

The partial effects of the main predictors in the model are illustrated in Figure 2.6. The upper left panel graphs the observed proportions of reduced forms for men and women: Men reduce more often than women. This may be due to the higher speech rate of men compared to women (Verhoeven et al., 2004). The upper right panel illustrates that words were less likely to be reduced in sentence final position. The lengthening of words in phrase-final position has been

Table 2.1: Coefficients in the logistic regression model (with dummy contrast coding) for the suffix reduction data. The intercept represents the log odds ratio of non-reduced to reduced words for Dutch highly educated women. The coefficients show the change in the log odds ratio accompanying the change from, e.g., women to men. See also Figure 2.6.

	Coefficient	Wald Z	p
Intercept	1.94	3.84	0.0001
Country: Flanders	0.58	0.87	0.3862
Sex: Male	−0.49	−2.67	0.0075
Position: Non-Final	−0.78	−3.10	0.0019
Education: not High	0.32	1.36	0.1726
Mutual Information	−0.11	−2.73	0.0064
Country: Flanders, & Education: Not High	−0.94	−2.47	0.0136

found to be a parsing cue for the listener (e.g., Scott, 1982). Reducing words in *-lijk* in phrase-final position would result in the absence of a useful perceptual cue, and apparently is avoided. The interaction of Country by Education is illustrated in the lower left panel. The factor Education is predictive only for Flanders: Flemish speakers with a high education level reduce less than non-highly educated Flemish speakers. As can be seen in Table 2.1, the coefficient for education (0.32), which describes the situation for the Netherlands, was not significant ($p = 0.17$). The lower right panel shows that reduced word forms had a higher mutual information. When words have a reduced information load, their forms can be less distinct as well.

The interaction of Country by Word is summarized in Figure 2.7. The horizontal axis indicates the adjustment that has to be made to express the amount of reduction of a given word in the Netherlands in relation to the amount of reduction of the word that is least often reduced, namely *moeilijk* ('difficult'), in the Netherlands. The vertical axis shows the adjustments for the amount of reduction of these words in Flanders, again compared to the amount of reduction of *moeilijk* in the Netherlands. The more negative the value on the axes, the greater the likelihood of reduction. Thus, the words in the upper left are relatively often reduced in the Netherlands, but are relatively seldom reduced in Flanders.

Note that the words *dadelijk* ('immediately'), *uiteindelijk* ('finally') and *tuurlijk* ('of course') differ in their behavior from the other words which are clustered in the lower right of the plot. Nevertheless, the interaction of Word by Country is still significant after the removal of these words. When the word *natuurlijk* is additionally removed, the interaction of Word by Country disappears. So, the behavior of the remaining 10 words is approximately the same in both countries.

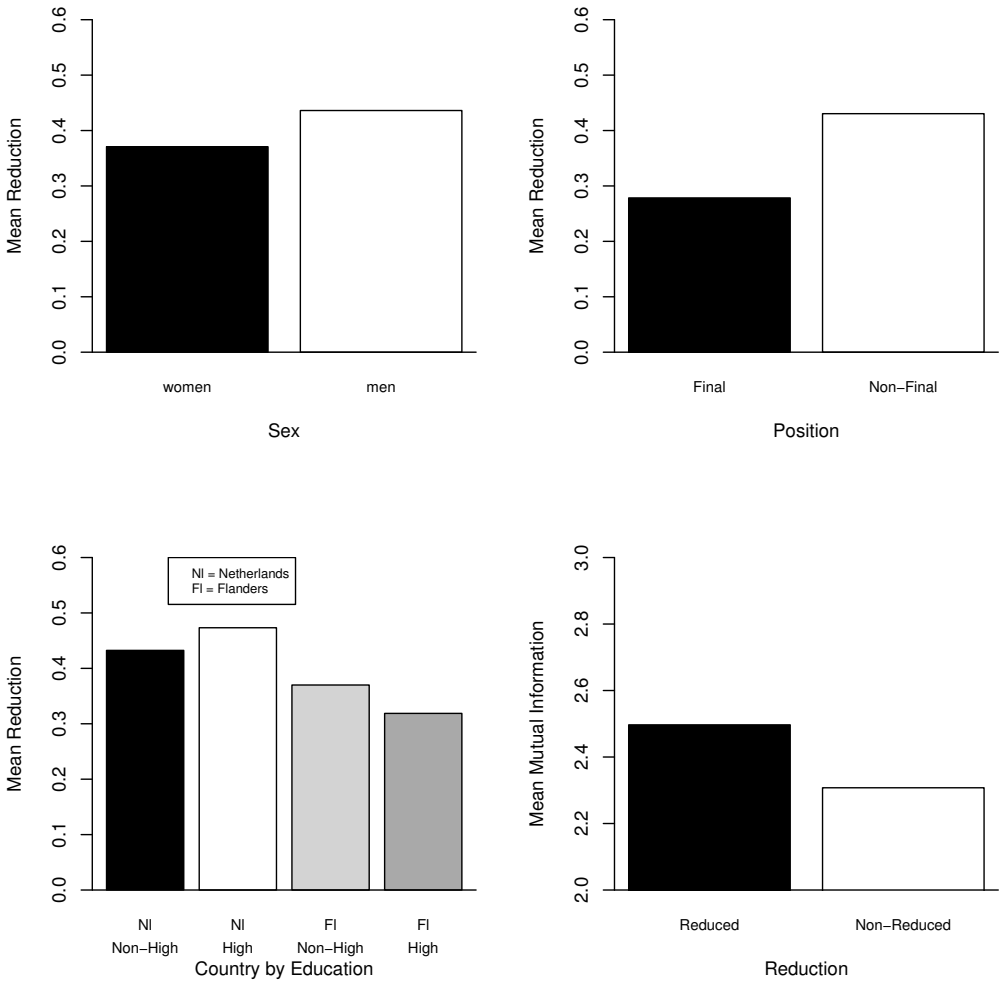


Figure 2.6: Observed proportion of reduced forms for 14 high-frequency words in *-lijk* broken down for Sex, for Position, and for both Country and Education. The lower right panel plots the mean Mutual Information for the reduced and unreduced forms of 14 high-frequency words in *-lijk*.

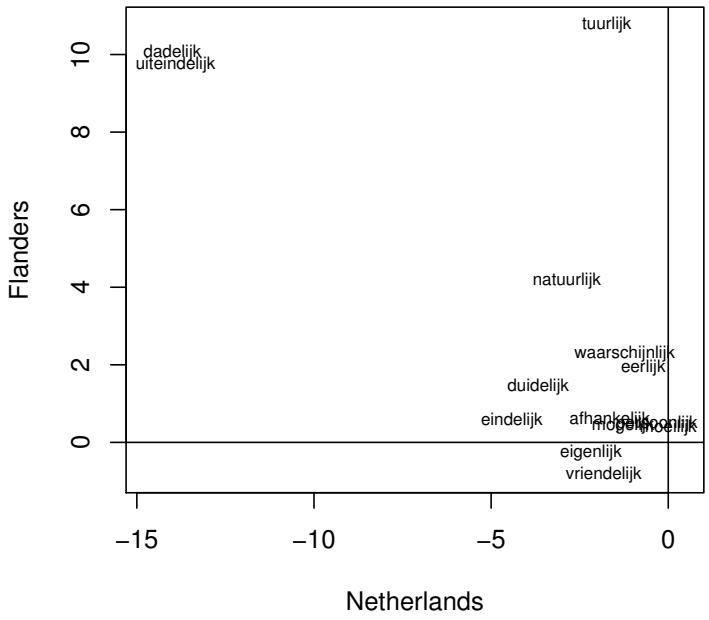


Figure 2.7: By Word adjustments for Country in a logistic regression model for 14 high-frequency words in *-lijk*.

Recall that these analyses are based on the log odds ratio of forms without reduction to forms with medium or high reduction. A sub-analysis of the eight words exhibiting three levels of reduction using a proportional odds model (Harrell, 2001) revealed a very similar pattern of results. We also ran an analysis in which we contrasted No or Medium Reduction with High Reduction. In this analysis, Sex and Mutual Information were no longer significant, while the other predictors were retained. This suggests that the effects of Sex and Mutual Information are mainly determined by differences between No and Medium Reduction.

Next, we analyzed the reduction of the vowel in the unstressed, word initial syllable, using the same 946 words we selected to explore variation in the reduction of the suffix. This kind of reduction occurred only in the three words *natuurlijk* ('of course'), *persoonlijk* ('personal'), and *waarschijnlijk* ('probably'). For *natuurlijk* we distinguished /nɑ/ and /na/ from /nə/ and /n/, for *persoonlijk* we distinguished /pə/ from /p/, and for *waarschijnlijk* we distinguished

/ʊɑ/ and /ʊa/ from /ʊə/ and /ʊ/. We used the same procedure as before: our point of departure was the abovementioned broad phonological transcription. This transcription was checked by an independent judge, who disagreed in 11 of the 194 cases. For these 11 tokens, a third judge decided their category assignment.

A binary logistic regression model fit to the reduction of the target words *natuurlijk*, *persoonlijk* and *waarschijnlijk* (using a stepwise model selection procedure) revealed a significant effect for Country ($X^2_{(1)} = 39.59$, $p < 0.0001$). Dutch speakers reduced the vowel in the unstressed word-initial syllable more often than Flemish speakers. The generalized R^2 index for this model was 0.232, and Somer's D_{xy} was 0.428 .

2.5 Conclusions

The aim of the present study was to explore the variation in the use of words in *-lijk* in both written and spoken Dutch. First, we investigated variation in written Dutch with respect to the country in which a text is written, and the text's register. Second, we explored spoken Dutch with respect to the speaker's country, level of education, and sex. For spoken Dutch, we investigated in more detail to what degree these socio-geographic factors codetermine the extent to which words in *-lijk* are acoustically reduced.

The methodology that we used for this investigation was inspired by prior stylometric studies on authorial variation (Burrows, 1992a) and studies on register variation (Biber, 1988) which used exploratory multivariate methods such as principal components analysis and factor analysis. We combined insights from these fields with insights from studies investigating the socio-geographical and socio-phonetic forces operating in language (see, e.g., Hay & Sudbury, 2005) in order to increase our understanding of variation in Dutch. Consequently, our study addresses an aggregation level (that of different social groups) that is intermediate between stylometry and authorship attribution (where the linguistic habits of individual authors are of central interest) and register variation (which typically studies texts sampled from a broad range of genres in spoken and written discourse).

Without denying the great value of principal components analysis, factor analysis, and correspondence analysis, we pursued a complementary approach using analysis of variance and covariance of lexical frequencies in factorially contrasted subcorpora. This methodology, which is tailored to our aim of studying the effect of socio-geographic factors on lexical variation, offers several advantages. One such advantage is that it becomes possible to test the significance of the design factors and their interactions with the individual words, without losing the possibility of visualization. Another advantage is that this methodology allows for the possibility of taking covariates into account. Finally, an advantage in relation to standard sociolinguistic practice in which individual controlled variables are studied in isolation, our approach makes it possible to

consider a great many potential carriers of sociolinguistic variation simultaneously. This allowed us to trace correlational structure between heterogeneous variables that otherwise remains invisible.

We first studied the variation in the frequency of use of words in *-lijk* in a corpus of Dutch newspapers. We selected all occurrences of 80 high-frequency words in *-lijk* from seven newspapers using a 2 by 3 factorial design. We distinguished between Flemish and Dutch newspapers (Country) and contrasted Quality newspapers (aiming at a more educated readership), National newspapers, and Regional newspapers (Register). In parallel, we conducted a study using the same design based on the 80 most common word types, following Burrows (1986, 1987, 1992ab, 1993ab).

This parallel study was motivated by the hypothesis that variation in the use of *-lijk* is unlikely to be isolated and encapsulated from other dimensions of variation in speech and writing (see Biber, 1988, 1995, for the many correlations into which grammatical markers enter). In order to properly understand the unique contribution of variation in *-lijk* to the linguistic profile of different groups of speakers, we needed a benchmark. Such a benchmark was provided by the covariance structure among the most common words, which tap into the syntactic habits of speakers, and therefore provide a shortcut to the more refined but also far more labor-intensive methods developed by Biber, which are feasible only for well-annotated corpora.

In both analyses, we observed significant and remarkably similar geographic and register differentiation. Apparently, high-frequency words in *-lijk* have a stylometric discriminatory potential that mirrors the well-established stylometric sensitivity of the most common words. Given that words in unproductive *-lijk* constitute a closed-class of words, this is a first way in which high-frequency words in *-lijk* have become to resemble the most common words, which mainly comprise closed-class function words such as conjunctions, pronouns, prepositions, and determiners.

Next, we explored the variation in frequency of use of words in *-lijk* in spoken Dutch. We selected 32 high-frequency words in *-lijk* from the subcorpora of spontaneous face-to-face conversations and telephone dialogues in the CGN, using a factorial design in which we contrasted speakers from Flanders with speakers from the Netherlands, men with women, and highly educated with less educated speakers. As before, we carried out a parallel study using the most common words. This time, we observed a marked difference between the most common words and the words in *-lijk*. Speakers with a higher education level tended to use words in *-lijk* more often. For the Netherlands (but not for Flanders), this mirrors the finding that the quality newspaper made more intensive use of this suffix as well. The analysis of the most common words, by contrast, suggested that men made less use of the most common words compared to women, suggesting the possibility of a slightly higher information density (carried by less frequent closed-class words or even by full-fledged content words) for men. In other words, the comparison with the benchmark for

grammatical variation revealed that in spoken Dutch, unlike in written Dutch, words in *-lijk* tap into an independent source of variation. Furthermore, we also observed significant differences in how individual most common words as well as individual words in *-lijk* were used by men and women in the two countries as a function of their education level.

Finally, we investigated the socio-geographic variation in the degrees of reduction of words in *-lijk*. This kind of research has become possible only recently, thanks to the development of large speech corpora with not only orthographic transcriptions but also the acoustic signal. Corpora such as the corpus of New Zealand English (Schreier et al., 2003; Gordon et al., forthcoming; Hay and Sudbury, 2005) and also the corpus of spoken Dutch offer the possibility of detailed analyses of the variation in acoustic forms across sociolinguistic and stylistic dimensions. The corpus of spoken Dutch was just large enough to allow us to retain our factorial methodology, although it left us with only 14 words ending in *-lijk* (evidencing reduction) that occurred sufficiently often in the different subcorpora defined by crossing Country, Sex, and Education. Two transcribers classified the degree of reduction for a total of 946 tokens of these 14 words. We considered two kinds of reduction, one primarily affecting the suffix, the other affecting the vowel in the word initial syllable. Both analyses showed that in Flanders speakers reduce less than in the Netherlands, which ties in with the more formal status of standard Dutch in Flanders. The reduction involving the suffix was more prominent for men compared to women. Moreover, highly educated Flemish speakers used fewer reduced forms than did less highly educated Flemish speakers. Finally, there were significant differences in the extent to which individual words underwent reduction that we could trace back to the speaker's home country. For instance, *dadelijk* ('deedly', i.e., 'immediately') and *uiteindelijk* ('end-ly', i.e., 'finally') are words that undergo reduction more often in the Netherlands than in Flanders. The degree of reduction is possibly influenced by speech rate. The higher the speech rate is, the more often reduction occurs. This assumption is strengthened by previous research in which, comparable to our results for reduction, it appeared that Dutch men have the highest speech rate, while Flemish women have the lowest (Verhoeven et al., 2004).

In addition to these socio-geographic factors, the degree of reduction was significantly codetermined by two linguistic factors: the word's position in the sentence, and the extent to which the word is predictable from its context. With respect to the word's position in the sentence, we found that words in *-lijk* that occurred in sentence-final position revealed little reduction. This is as expected given that words in sentence final position are often lengthened (e.g., Fougeron and Keating, 1997; Cambier-Langeveld, 2000; Pluymaekers et al., submitted).

We used the mutual information measure to gauge contextual predictivity. Words in *-lijk* with a high mutual information (Manning and Schutze, 1999), i.e., that exhibited a high degree of predictability from the preceding word,

revealed more reduction. As the information load of a word in *-lijk* decreases, speakers fall back gestural scores that require less articulatory effort in production (see cf. Bybee, 2005).

The overall pattern in our data suggests that reduced high-frequency forms in *-lijk*, such as monosyllabic [tyk] (for *natuurlijk*, 'of course'), [mok] (for *mogelijk*, 'possible') and [ɛɪk] (for *eigenlijk*, 'actually') are becoming more similar to the most common words, not only in that they are markers of register and of socio-geographic origin, as observed above, but also in their loss of morphological structure, as witnessed by their lack of semantic compositionality and the erosion (Heine and Kuteva, 2005) of their phonological form.

Interestingly, the position in the sentence and mutual information are contextual predictors that did not interact with the socio-geographic variables. This is reminiscent of the finding of Bresnan et al. (2005) that the formal syntactic and semantic properties governing the dative alternation in English do not change across modality (spoken versus written English), verb sense, and speaker. This suggests that there are robust fundamental linguistic principles that operate in the same way across register and different socio-geographic speech communities. Possibly, phrase-final lengthening and information load belong to the set of these fundamental principles. Further research addressing acoustic reduction for other kinds of complex words is required here.

What the present results clearly show is that for a full explanation of acoustic reduction socio-geographic factors need to be taken into account. Although articulatory explanations (such as offered in Browman and Goldstein, 1992; Ernestus, 2000) increase our insight in the path of acoustic erosion, they do not predict when speakers actually use reduced forms and when they stick with the unreduced forms. We have shown that some headway in predicting degrees of reduction can be made by taking socio-geographic factors and contextual linguistic factors into account.

For [mok], [ɛɪk] and [tyk] a large series of different forms exist side by side. The unreduced long, morphologically complex, but semantically opaque forms are predominant in the written language, and shape modern speakers' awareness of these words. The reduced, monosyllabic forms are typically found in spontaneous spoken Dutch, without speakers realizing that what they actually say diverges from the written norms (Kemps et al., 2004). Reduced forms challenge models of speech comprehension, which are based on the assumption that words have a single canonical form (Norris, 1994). The present results suggest that listeners might be sensitive to the probability of reduced forms conditional on the socio-geographic and linguistic context in which a word in *-lijk* is uttered, and use this sensitivity to optimize comprehension.

It is well known that morphological rules can cease to be productive. For some affixes, the change from productive to unproductive takes place in a relative short time span, while for others, the change is more gradual (Anshen and Aronoff, 1997, 1999).

However, one of the questions in productivity research is whether there

is an absolute distinction between productive and unproductive affixes. Many morphologists believe there is such a distinction (e.g., Schultink, 1961; Anshen and Aronoff, 1999; Bauer, 2001), but there is evidence to the contrary. Baayen (2003) discusses well-formed and contextually natural neologisms in English *-th* (e.g., *coolth*), and a search on the web shows that neologisms in *-lijk* are likewise in use in Dutch. Here are two examples, and many more can be found.

> *Beschrijf jezelf in 5 woorden: lief agressief aardig* **bazelijk** *en gek*
> (Describe yourself in 5 words: sweet, aggressive, nice, bossy, and mad)
> `www.dreamcommunity.nl/?id=109&account=jEHA2yBJ` (May 2005)

> *. . . 's ochtends ineens kreeg ik erge spierpijn en werd ik wazig in m'n hoofd,* **duizelijk** *en zweverig, buikkrampen en een misselijk gevoel . . .*
> (in the morning I suddenly developed muscular pain, I became drowsy in the head, dizzy, woolly, stomach cramps, and I felt sick . . .)
> `www.degrotegriepmeting.nl/test/public/index.php?thissection_id=3& request=1150&r=1` (May 2005)

Even though neologisms in *-lijk* have very low probabilities of being coined, there are sufficient numbers of words in *-lijk* in the language for speakers to be able to occasionally generalize *-lijk* to new words. The second example shows that even the blocking force of existing synonyms (*duizelijk* replaces the standard form *duize-lig*) may not prevent new words to be used effectively. From this perspective, the erosion of high-frequency words in *-lijk* is interesting, as this erosion results in words that no longer contribute to the formal similarities that underly this residual productivity of *-lijk*. We expect that as more high-frequency words undergo this process of erosion and become effectively monomorphemic, the residual productivity of *-lijk* will decrease even further. Independent evidence that we are indeed observing a still ongoing process of decreasing productivity and language change is provided by the finding reported by Pluymakers et al. (2005) that younger speakers tend to reduce words in *-lijk* to a greater extent then older speakers, and the present finding that reduction is more prominent in the Netherlands than in Flanders (where standard Dutch is used predominantly in more formal contexts).

References

Anshen, F. and M. Aronoff, 1997. Morphology in real time. In G. E. Booij and J. van Marle, eds., *Yearbook of Morphology*. Kluwer Academic Publishers, Dordrecht, 9–12

Anshen, F. and M. Aronoff, 1999. Using dictionaries to study the mental lexicon. *Brain and Language*, 68: 16–26

Baayen, R. H., 1994. Derivational productivity and text typology. *Journal of Quantitative Linguistics*, 1: 16–34

Baayen, R. H., 2003. Probabilistic approaches to morphology. In R. Bod, J. Hay and S. Jannedy, eds., *Probabilistic linguistics*. The MIT Press, 229–287

Baayen, R. H., 2004. Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers*: 1–45

Baayen, R. H., H. van Halteren, A. Neijt and F. J. Tweedie, 2002. An experiment in authorship attribution. In *Proceedings of JADT 2002*. Université de Rennes, St. Malo, 29–37

Baayen, R. H., H. van Halteren and F. Tweedie, 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11: 121–131

Bauer, L., 2001. *Morphological productivity*. Cambridge University Press, Cambridge

Biber, D., 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge

Biber, D., 1995. *Dimensions of Register Variation*. Cambridge University Press, Cambridge

Browman, C. P. and L. Goldstein, 1995. Dynamics and articulatory phonology. In T. Van Gelder and R. F. Port, eds., *Mind as motion*. MIT Press, Cambridge, Massachusetts, 175–193

Burrows, J. F., 1986. Modal verbs and moral principles: An aspect of Jane Austen's style. *Literary and Linguistic Computing*, 1: 9–23

Burrows, J. F., 1987. Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, 2: 61–70

Burrows, J. F., 1992a. Computers and the study of literature. In C. S. Butler, ed., *Computers and Written Texts*. Blackwell, Oxford, 167–204

Burrows, J. F., 1992b. Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7: 91–109

Burrows, J. F., 1993a. Noisy signals? Or signals in the noise? In *ACH-ALLC Conference Abstracts*. Georgetown, 21–23

Burrows, J. F., 1993b. Tiptoeing into the infinite: Testing for evidence of national differences in the language of English narrative. In S. Hockey and N. Ide, eds., *Research in Humanities Computing '92*. Oxford University Press, London

Bybee, J. L., 2001. *Phonology and language use*. Cambridge University Press, Cambridge

Bybee, J. L., 2005. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *manuscript submitted for publication*

Cambier-Langeveld, T., 2000. *Temporal marking of accents and boundaries*. LOT, Amsterdam

Chatterjee, S., A. S. Hadi and B. Price, 2000. *Regression analysis by example*. John Wiley & Sons, New York

Church, K. W. and P. Hanks, 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16: 22–29

Ernestus, M., 2000. *Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface*. LOT, Utrecht

Ernestus, M., R. H. Baayen and R. Schreuder, 2002. The recognition of reduced word forms. *Brain and Language*, 81: 162–173

Fougeron, C. and P. Keating, 1997. Articulatory strengthening at the edges of prosodic domains. *Journal of the Acoustical Society of America*, 101(6): 3728–3740

Gordon, E., M. Maclagan and J. Hay, forthcoming. The ONZE Corpus. In J. C. Beal, K. P. Corrigan and H. Moisl, eds., *Models and methods in handling of unconventional digital corpora*, volume 2. Palgrave

Gregory, M. L., W. D. Raymond, A. Bell, E. Fosler-Lussier and D. Jurafsky, 1999. The effects of collocational strength and contextual predictability in lexical production. *CLS*, 35: 151–166

Grondelaers, S., K. Deygers, H. van Aken, V. van den Heede and D. Speelman, 2000. Het CONDIV-corpus geschreven Nederlands. *Nederlandse Taalkunde*, 5: 356–363

Harrell, F. E., 2001. *Regression modeling strategies*. Springer, Berlin

Hay, J. and A. Sudbury, 2005. How rhoticity became/r/-sandhi. *Language*, 81 (4): 799–823

Heine, B. and T. Kuteva, 2005. *Language Contact and Grammatical Change*. Cambridge University Press, Cambridge

Holmes, D. I., 1994. Authorship attribution. *Computers and the Humanities*, 28 (2): 87–106

Johnson, K., 2004. Massive reduction in conversational American English. In *Spontaneous speech: data and analysis. Proceedings of the 1st session of the 10th international symposium*. The National International Institute for Japanese Language, Tokyo, Japan, 29–54

Jurafsky, D., A. Bell, M. Gregory and W. Raymond, 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee and P. Hopper, eds., *Frequency and the emergence of linguistic structure*. John Benjamins, Amsterdam, 229–254

Kemps, R., M. Ernestus, R. Schreuder and R. H. Baayen, 2004. Processing reduced word forms: The suffix restoration effect. *Brain and Language*, 90: 117–127

Lebart, L., A. Salem and L. Berry, 1998. *Exploring Textual Data*. Kluwer, Dordrecht

Manning, C. D. and H. Schutze, 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge

Norris, D. G., 1994. Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52: 189–234

Oostdijk, N. H. J., 2002. The Design of the Spoken Dutch Corpus. In P. Peters, P. Collins and A. Smith, eds., *New Frontiers of Corpus Research*. Rodopi, Amsterdam, 105–112

Pinheiro, J. C. and D. M. Bates, 2000. *Mixed-effects models in S and S-PLUS*. Statistics and Computing, Springer, New York

Plag, I., C. Dalton-Puffer and R. H. Baayen, 1999. Productivity and register. *Journal of English Language and Linguistics*, 3: 209–228

Pluymakers, M., M. Ernestus and R. H. Baayen, 2005. Lexical frequency and acoustic reduction in spoken dutch. *Journal of the Acoustical Society of America*, 18: 2561–2569

Quené, H. and H. van den Bergh, 2004. On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Communication*, 43: 103–121

Schreier, D., E. Gordon, J. Hay and M. Maclagan, 2003. The regional and linguistic dimension of /hw/ maintenance and loss in early 20^{th} century New Zealand English. *English World-Wide*, 24 (2): 245–270

Schultink, H., 1961. Produktiviteit als morfologisch fenomeen. *Forum der Letteren*, 2: 110–125

Scott, D. R., 1982. Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America*, 71: 996–1007

Van Marle, J., 1988. Betekenis als factor bij productiviteitsverandering. *Spektator*, 17: 341–359

Verhoeven, J., G. de Pauw and H. Kloots, 2004. Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language and Speech*, 47 (3): 297–308

Zipf, G. K., 1935. *The Psycho-Biology of Language.* Houghton Mifflin, Boston

Appendix A

Selected words ending in the suffix *-lijk* from the newspaper corpus (CONDIV):
mogelijk; duidelijk; natuurlijk; eigenlijk; uiteindelijk; moeilijk waarschijnlijk; namelijk; onmiddellijk; eindelijk; verantwoordelijk; aanvankelijk; gemakkelijk; onmogelijk; persoonlijk; makkelijk; degelijk; vermoedelijk; noodzakelijk; behoorlijk; gevaarlijk; tijdelijk; voornamelijk; afhankelijk; kennelijk; eerlijk; letterlijk; aanzienlijk; werkelijk; koninklijk; opmerkelijk; redelijk; respectievelijk; gezamenlijk; wettelijk; onduidelijk; hopelijk; onafhankelijk; gedeeltelijk; daadwerkelijk; wetenschappelijk; gerechtelijk; dergelijk; toegankelijk; oorspronkelijk; landelijk; gemeentelijk; maatschappelijk; aantrekkelijk; menselijk; nadrukkelijk; stedelijk; onvermijdelijk; openlijk; verschrikkelijk; heerlijk; uitzonderlijk; geleidelijk; voorwaardelijk; tamelijk; ongelooflijk; vriendelijk; dodelijk; pijnlijk; vreselijk; herhaaldelijk; plaatselijk; vrolijk; belachelijk; schriftelijk; hoofdzakelijk; gebruikelijk; uitdrukkelijk; onbegrijpelijk; gewoonlijk; begrijpelijk; inhoudelijk; onwaarschijnlijk; feitelijk; verwonderlijk.

Most common words from the newspaper corpus (CONDIV):
de; van; het; een; in; en; dat; op; is; te; voor; met; zijn; die; niet; aan; er; maar; ik; om; als; ook; hij; bij; uit; nog; door; naar; heeft; ze; was; dan; over; tot; jaar; worden; we; of; al; wordt; meer; hebben; je; zich; geen; werd; kan; dit; zo; wat; hun; na; wel; nu; moet; tegen; twee; deze; kunnen; haar; veel; had; uur; eerste; zou; zal; nieuwe; onder; moeten; daar; andere; wil; volgens; gaat; mijn; toch; mensen; waar; gaan; zegt.

Selected words ending in the suffix *-lijk* from the corpus of spoken Dutch (CGN):
eigenlijk; natuurlijk; waarschijnlijk; tuurlijk; moeilijk; uiteindelijk; redelijk; makkelijk; duidelijk; mogelijk; verschrikkelijk; eerlijk; dadelijk; gemakkelijk; belachelijk; onmiddellijk; ongelooflijk; eindelijk; tamelijk; lelijk; persoonlijk; gevaarlijk; afhankelijk; vriendelijk; vrolijk; ongelofelijk; hopelijk; voornamelijk; degelijk; oorspronkelijk; feitelijk; pijnlijk.

Most common words from the corpus of spoken Dutch (CGN):
aan; ah; al; als; ben; bij; daar; dan; da's; dat; de; denk; die; doen; d'r; dus; echt; een; eigenlijk; en; er; gaan; gaat; ge; gewoon; goed; had; hè; heb; hebben; heeft; heel; het; hij; hoe; ie; ik; in; is; ja; je; 'k; kan; keer; maar; meer; met; mij; mmm; moet; naar; nee; niet; nog; nu; of; oh; om; ook; op; 't; te; toch; toen; uh; uhm; van; veel; voor; want; was; wat; we; weer; weet; wel; ze; zeg; zijn; zo.

Appendix B

Table 2.2: Values of the coefficients as visualized in Figure 2.2

Word	Netherl.	Flanders	Quality	National	Regional
aantrekkelijk	0.14	−0.49	0.04	−0.02	−0.17
aanvankelijk	0.21	−0.30	0.01	0.02	−0.15
aanzienlijk	0.23	−0.48	0.02	0.09	−0.25
afhankelijk	0.22	−0.45	0.05	−0.12	−0.07
begrijpelijk	0.09	−0.56	0.04	−0.05	−0.15
behoorlijk	−0.07	−0.13	−0.05	0.06	−0.14
belachelijk	−0.19	−0.25	−0.05	−0.02	−0.08
daadwerkelijk	0.37	−0.70	−0.01	0.01	−0.15
degelijk	−0.11	−0.06	−0.03	−0.01	−0.09
dergelijk	0.01	−0.34	−0.04	−0.01	−0.10
dodelijk	−0.17	−0.23	−0.06	−0.01	−0.08
duidelijk	0.10	0.09	−0.05	0.04	−0.09
eerlijk	−0.01	−0.22	−0.04	−0.02	−0.08
eigenlijk	0.25	−0.14	−0.05	−0.03	−0.02
eindelijk	0.04	−0.11	−0.16	0.13	−0.09
feitelijk	0.30	−0.81	−0.02	0.03	−0.18
gebruikelijk	0.32	−0.80	0.05	−0.08	−0.13
gedeeltelijk	−0.10	−0.22	0.03	−0.12	−0.06
geleidelijk	−0.24	−0.16	0.11	−0.05	−0.21
gemakkelijk	0.06	−0.16	0.05	0.02	−0.19
gemeentelijk	−0.52	0.13	−0.13	−0.01	−0.02
gerechtelijk	−0.66	0.28	0.07	−0.08	−0.15
gevaarlijk	−0.11	−0.10	−0.03	0.01	−0.11
gewoonlijk	−0.20	−0.27	0.00	−0.03	−0.13
gezamenlijk	0.32	−0.62	0.06	−0.03	−0.17
heerlijk	0.19	−0.58	−0.04	−0.02	−0.09
herhaaldelijk	−0.17	−0.25	0.02	−0.06	−0.12
hoofdzakelijk	−0.40	−0.06	0.05	−0.03	−0.18
hopelijk	−0.55	0.14	−0.07	−0.03	−0.06
inhoudelijk	−0.03	−0.43	0.04	−0.08	−0.12
kennelijk	0.68	−0.95	0.07	−0.03	−0.18
koninklijk	−0.25	−0.03	0.17	−0.12	−0.19
landelijk	0.73	−1.14	0.00	−0.13	−0.03
letterlijk	0.04	−0.28	−0.03	−0.06	−0.05
maatschappelijk	−0.01	−0.35	0.12	−0.20	−0.07
makkelijk	−0.03	−0.14	−0.02	−0.04	−0.07
menselijk	−0.09	−0.28	0.08	−0.12	−0.12
moeilijk	0.01	0.08	−0.01	0.02	−0.11

Table 2.2: continued

Word	Netherl.	Flanders	Quality	National	Regional
mogelijk	0.25	0.00	−0.03	0.02	−0.08
nadrukkelijk	0.30	−0.68	−0.01	−0.05	−0.10
namelijk	0.32	−0.35	−0.13	0.12	−0.11
natuurlijk	0.24	−0.06	−0.09	0.05	−0.05
noodzakelijk	0.01	−0.19	0.05	−0.05	−0.13
onafhankelijk	0.18	−0.50	0.03	−0.02	−0.16
onbegrijpelijk	−0.10	−0.36	−0.04	0.04	−0.16
onduidelijk	0.28	−0.59	0.06	−0.06	−0.14
ongelooflijk	−0.13	−0.27	−0.10	0.02	−0.07
onmiddellijk	−0.34	0.29	−0.05	0.06	−0.12
onmogelijk	−0.07	−0.06	−0.01	−0.02	−0.10
onvermijdelijk	−0.02	−0.36	0.12	−0.15	−0.12
onwaarschijnlijk	−0.09	−0.40	0.03	0.00	−0.19
oorspronkelijk	−0.22	−0.13	−0.01	−0.10	−0.05
openlijk	0.01	−0.39	0.09	−0.07	−0.18
opmerkelijk	0.03	−0.29	−0.01	0.01	−0.14
persoonlijk	0.16	−0.31	0.00	−0.02	−0.11
pijnlijk	−0.15	−0.25	0.02	−0.01	−0.17
plaatselijk	−0.64	0.17	0.01	−0.09	−0.08
redelijk	0.38	−0.66	−0.06	0.05	−0.14
respectievelijk	0.04	−0.32	−0.02	0.06	−0.18
schriftelijk	−0.08	−0.36	−0.04	−0.06	−0.06
stedelijk	−0.53	0.12	−0.05	−0.02	−0.09
tamelijk	0.46	−0.90	0.08	−0.03	−0.21
tijdelijk	0.03	−0.25	−0.02	0.00	−0.11
toegankelijk	−0.23	−0.11	0.00	−0.08	−0.07
uitdrukkelijk	−0.47	0.00	0.02	−0.08	−0.11
uiteindelijk	0.16	−0.07	−0.13	0.08	−0.05
uitzonderlijk	−0.40	0.01	0.10	−0.12	−0.14
verantwoordelijk	0.24	−0.33	0.06	−0.08	−0.11
vermoedelijk	−0.21	0.02	−0.05	0.04	−0.13
verschrikkelijk	−0.08	−0.30	−0.04	−0.03	−0.08
verwonderlijk	−0.22	−0.29	0.01	0.03	−0.20
voornamelijk	0.23	−0.46	0.02	−0.08	−0.07
voorwaardelijk	−0.03	−0.42	−0.11	0.04	−0.09
vreselijk	0.20	−0.63	−0.03	−0.07	−0.05
vriendelijk	−0.14	−0.28	−0.03	−0.12	−0.01
vrolijk	0.10	−0.53	0.01	−0.03	−0.13
waarschijnlijk	0.24	−0.24	−0.02	−0.02	−0.07
werkelijk	0.08	−0.33	0.09	−0.07	−0.16
wetenschappelijk	−0.16	−0.16	0.06	−0.08	−0.13
wettelijk	−0.07	−0.23	0.03	−0.06	−0.12

Table 2.3: Values of the coefficients as visualized in Figure 2.3

Word	Nl Quality	Adj. Flanders	Adj. National	Adj. Regional
aan	0.65	−0.11	0.02	0.01
al	−0.21	−0.10	0.08	0.07
alleen	−0.98	−0.33	−0.13	−0.06
als	0.66	−0.34	−0.22	−0.12
andere	−0.87	−0.23	−0.10	−0.04
bij	0.34	−0.21	0.01	0.01
daar	−0.99	−0.20	0.07	0.10
dan	0.08	−0.17	−0.04	−0.01
dat	1.47	−0.23	−0.14	−0.05
de	3.11	−0.14	−0.02	−0.03
deze	−0.58	−0.26	−0.07	−0.04
die	1.14	−0.22	−0.20	−0.11
dit	−0.43	−0.19	0.04	0.02
door	0.38	−0.35	−0.15	−0.11
drie	−1.28	−0.09	0.15	0.13
een	2.14	−0.15	−0.08	−0.07
eerste	−1.00	−0.02	0.19	0.10
en	1.90	−0.10	−0.08	−0.09
er	0.46	−0.08	0.07	0.06
gaan	−1.14	−0.24	0.05	0.08
gaat	−0.99	−0.27	−0.03	0.03
geen	−0.28	−0.16	−0.02	0.02
haar	−0.65	−0.21	−0.08	−0.02
had	−0.71	−0.24	−0.05	−0.02
hebben	−0.18	−0.25	−0.02	0.04
heeft	0.31	−0.37	−0.13	−0.02
het	2.26	−0.18	−0.08	−0.03
hij	0.55	−0.26	−0.14	−0.06
hun	−0.45	−0.15	−0.06	−0.03
ik	0.32	0.03	0.11	0.09
in	2.03	−0.17	−0.05	−0.04
is	1.40	−0.27	−0.12	−0.04
je	−0.22	−0.26	−0.07	0.05
kan	−0.46	−0.11	0.02	0.03
kunnen	−0.67	−0.16	0.04	0.04
maar	0.57	−0.16	−0.05	−0.03
meer	−0.17	−0.18	−0.03	0.01
met	1.11	−0.16	0.00	−0.02
mijn	−1.30	0.00	0.26	0.16
moet	−0.57	−0.17	0.02	0.05
moeten	−0.98	−0.16	0.06	0.07

Table 2.3: continued

Word	Nl Quality	Adj. Flanders	Adj. National	Adj. Regional
na	−0.58	−0.11	0.12	0.05
naar	0.13	−0.15	0.05	0.06
niet	0.92	−0.17	−0.07	−0.01
nieuwe	−0.91	−0.18	0.02	0.04
nog	0.06	−0.07	0.15	0.11
nu	−0.54	−0.13	0.06	0.04
of	0.06	−0.25	−0.23	−0.14
om	0.38	−0.08	0.07	0.05
omdat	−1.10	−0.24	−0.02	0.04
onder	−0.78	−0.31	−0.11	−0.06
ook	0.43	−0.17	0.01	0.02
op	1.28	−0.10	0.06	0.01
over	0.26	−0.34	−0.26	−0.16
te	1.26	−0.24	−0.08	−0.06
tegen	−0.67	−0.09	0.12	0.04
toch	−1.23	−0.07	0.13	0.11
tot	−0.07	−0.19	−0.03	−0.02
twee	−0.73	−0.08	0.15	0.11
uit	0.16	−0.11	0.06	0.05
van	2.46	−0.17	−0.14	−0.13
veel	−0.64	−0.31	−0.09	−0.03
volgens	−0.90	−0.31	−0.07	0.00
voor	1.13	−0.11	0.00	−0.01
waar	−1.03	−0.29	−0.07	0.00
was	0.04	−0.14	0.04	0.02
wat	−0.42	−0.16	−0.08	−0.03
we	−0.53	0.06	0.28	0.26
wel	−0.52	−0.20	0.04	0.05
werd	−0.69	0.09	0.27	0.13
wil	−1.00	−0.18	0.05	0.09
worden	−0.02	−0.24	−0.06	−0.02
wordt	−0.17	−0.20	−0.02	0.01
zal	−0.87	−0.14	0.13	0.09
ze	0.22	−0.18	−0.04	0.02
zegt	−1.14	−0.17	0.11	0.12
zich	−0.13	−0.27	−0.14	−0.08
zijn	1.20	−0.23	−0.11	−0.07
zo	−0.33	−0.20	−0.11	−0.10
zou	−0.73	−0.24	−0.10	−0.05

Table 2.4: Values of the coefficients as visualized in Figure 2.4

Word	Word Freq.	Adj. Flanders	Adj. Men	Adj. Edu-Non-High
afhankelijk	−0.96	0.07	0.21	−0.29
belachelijk	−0.01	−0.05	0.03	−0.11
dadelijk	0.55	−1.47	−0.38	0.34
degelijk	−1.83	0.45	0.09	−0.15
duidelijk	0.64	−0.29	0.27	−0.37
eerlijk	0.32	−0.53	0.06	−0.14
eigenlijk	3.58	0.55	0.01	−0.21
eindelijk	−0.04	−0.29	−0.20	0.13
feitelijk	−2.57	1.23	0.30	−0.37
gemakkelijk	−1.24	2.17	0.16	−0.30
gevaarlijk	−0.92	0.23	0.00	−0.07
hopelijk	−1.80	1.15	0.14	−0.23
lelijk	−0.69	−0.12	−0.26	0.20
makkelijk	1.06	−1.95	−0.12	0.08
moeilijk	1.52	0.40	−0.05	−0.09
mogelijk	0.51	0.18	0.12	−0.23
natuurlijk	3.55	−0.42	0.11	−0.27
ongelofelijk	−1.55	0.51	0.40	−0.48
ongelooflijk	−0.93	1.14	0.50	−0.63
onmiddellijk	−1.75	2.32	0.54	−0.69
oorspronkelijk	−1.77	0.34	0.04	−0.10
persoonlijk	−0.93	0.47	0.09	−0.17
pijnlijk	−2.50	1.13	0.19	−0.26
redelijk	0.55	0.56	0.25	−0.38
tamelijk	−2.10	2.11	0.61	−0.74
tuurlijk	1.87	−0.03	−0.03	−0.11
uiteindelijk	0.81	0.92	0.09	−0.23
verschrikkelijk	0.59	0.01	−0.19	0.10
voornamelijk	−1.14	−0.43	0.11	−0.16
vriendelijk	−1.51	1.22	−0.35	0.27
vrolijk	−0.87	−1.39	−0.20	0.19
waarschijnlijk	1.77	0.82	0.09	−0.26

Table 2.5: Values of the coefficients as visualized in Figure 2.5

Word	Word Freq.	Adj. Flanders	Adj. Men	Adj. Edu-Non-High
aan	−0.0337	0.0086	0.0002	0.0012
ah	−0.1088	0.1307	−0.0010	0.0080
al	−0.0316	0.0157	−0.0059	0.0029
als	−0.0047	0.0025	0.0000	−0.0014
ben	−0.0535	−0.0138	−0.0039	0.0032

Table 2.5: continued

Word	Word Freq.	Adj. Flanders	Adj. Men	Adj. Edu-Non-High
bij	−0.0261	−0.0159	−0.0036	0.0023
daar	−0.0003	0.0130	0.0016	−0.0034
dan	0.0815	−0.0078	−0.0072	−0.0001
da's	−0.0454	0.0398	−0.0008	0.0021
dat	0.1386	0.0278	0.0022	−0.0025
de	0.0756	0.0012	0.0148	−0.0056
denk	−0.0519	−0.0004	−0.0031	−0.0012
die	0.0838	−0.0099	0.0083	−0.0033
doen	−0.0526	0.0067	−0.0065	0.0011
d'r	−0.0253	−0.0241	0.0031	−0.0014
dus	0.0313	−0.0050	0.0020	−0.0103
echt	−0.0198	−0.0193	−0.0097	−0.0051
een	0.0863	−0.0184	0.0105	−0.0083
eigenlijk	−0.0670	0.0296	0.0006	−0.0076
en	0.1246	0.0079	−0.0052	0.0011
er	−0.0386	0.0095	0.0024	−0.0051
gaan	−0.0547	0.0245	−0.0042	0.0022
gaat	−0.0646	0.0096	−0.0047	0.0031
ge	−0.1657	0.1785	−0.0025	0.0150
gewoon	−0.0111	−0.0563	−0.0078	−0.0045
goed	−0.0330	0.0036	−0.0035	0.0008
had	−0.0228	−0.0237	−0.0124	0.0032
hé	−0.0164	0.0660	0.0041	0.0044
heb	0.0122	−0.0203	−0.0032	0.0006
hebben	−0.0245	−0.0078	−0.0021	−0.0050
heeft	−0.0481	0.0062	−0.0072	−0.0021
heel	−0.0188	−0.0275	−0.0107	−0.0061
het	−0.0464	0.0694	0.0009	−0.0137
hij	−0.0317	0.0033	−0.0077	−0.0018
hoe	−0.0565	−0.0020	−0.0018	0.0029
ie	−0.0270	−0.0868	−0.0047	−0.0008
ik	0.1339	−0.0240	−0.0071	0.0019
in	0.0371	0.0056	0.0053	−0.0060
is	0.0777	0.0125	0.0032	−0.0040
ja	0.2067	0.0064	0.0027	−0.0052
je	0.1008	−0.1137	−0.0008	−0.0149
'k	0.0079	0.0422	−0.0160	0.0155
kan	−0.0313	−0.0183	−0.0061	−0.0008
keer	−0.0663	0.0166	−0.0054	0.0052
maar	0.0909	0.0012	−0.0056	0.0008
meer	−0.0550	0.0023	−0.0042	0.0007
met	0.0036	0.0035	−0.0003	0.0019

Table 2.5: continued

Word	Word Freq.	Adj. Flanders	Adj. Men	Adj. Edu-Non-High
mij	−0.0630	0.0077	−0.0027	−0.0005
mmm	−0.0534	−0.0161	−0.0012	−0.0024
moet	−0.0160	0.0039	−0.0037	0.0028
naar	−0.0346	0.0122	−0.0050	0.0044
nee	0.0426	−0.0250	0.0014	−0.0001
niet	0.0747	0.0084	−0.0096	0.0040
nog	0.0149	0.0068	−0.0035	0.0027
nu	−0.0677	0.0635	−0.0033	−0.0016
of	0.0163	−0.0080	0.0016	−0.0050
oh	0.0369	−0.1126	−0.0210	0.0043
om	−0.0430	0.0077	0.0011	−0.0069
ook	0.0656	−0.0242	−0.0089	−0.0027
op	0.0122	−0.0053	0.0036	−0.0028
’t	0.0970	−0.0114	−0.0033	0.0029
te	−0.0159	0.0130	0.0024	−0.0054
toch	−0.0176	−0.0047	−0.0075	0.0016
toen	−0.0225	−0.0653	−0.0123	0.0050
uh	0.1376	−0.0638	0.0230	−0.0106
uhm	−0.0478	0.0159	0.0034	−0.0114
van	0.0418	0.0132	0.0056	−0.0015
veel	−0.0625	0.0089	−0.0028	−0.0048
voor	−0.0187	0.0141	0.0005	−0.0004
want	0.0000	−0.0189	−0.0180	0.0030
was	0.0140	0.0045	−0.0052	0.0006
wat	0.0057	−0.0175	0.0055	−0.0015
we	0.0052	−0.0091	−0.0006	−0.0035
weer	−0.0360	−0.0485	−0.0060	0.0036
weet	−0.0172	−0.0021	−0.0126	0.0070
wel	0.0584	−0.0271	−0.0071	0.0002
ze	0.0341	0.0031	−0.0165	0.0081
zeg	−0.0528	−0.0055	−0.0152	0.0109
zijn	−0.0230	0.0296	0.0031	−0.0054
zo	0.0298	0.0210	−0.0110	0.0054

Table 2.6: Values of the coefficients as visualized in Figure 2.7

Word	Netherlands	Flanders
afhankelijk	−1.65	0.58
dadelijk	−13.99	10.05
duidelijk	−3.68	1.43
eerlijk	−0.70	1.93
eigenlijk	−2.18	−0.27
eindelijk	−4.42	0.57
moeilijk	0.00	0.38
mogelijk	−1.27	0.42
natuurlijk	−2.86	4.17
persoonlijk	−0.32	0.48
tuurlijk	−1.73	10.76
uiteindelijk	−13.91	9.74
vriendelijk	−1.82	−0.83
waarschijnlijk	−1.23	2.29

CHAPTER 3

Socio-geographic variation in morphological productivity in spoken Dutch: a comparison of statistical techniques[1]

Abstract

This study explores socio-geographic variation in morphological productivity in spoken Dutch. For 72 affixes, we extracted the hapax legomena from the Corpus of Spoken Dutch. We divided the corpus into 24 subcorpora defined by the speaker's country (Flanders versus the Netherlands), education level (High versus Non-High), sex (Women versus Men), and age (Young, Mid or Old). The large number of cells with zero counts for the affixes, and the substantial variation in the sizes of the subcorpora underlying the cell counts, posed a special challenge for the statistical analysis. We fitted three different kinds of models to our data: an ordinary least squares linear model with a transformation of the proportion of hapax legomena in the subcorpus as dependent variable, a linear mixed effects model with affix as random effect and the transformed proportions as dependent variable, and a generalized linear model with a binomial link which considered the hapax legomena as successes, and all remaining words as failures. The generalized linear model outperformed the others, in spite of the extremely small probabilities of success. We discuss why the generalized linear model is superior, and show how generalized linear models can be used to visualize by-affix variability in productivity.

Keywords: socio-geographic variation, generalized linear models, morphology, productivity, visualization.

[1]This study, co-authored by Roeland van Hout and Harald Baayen, is published under the same title in J.-M. Viprey, editor, *Actes des 8es journées internationales d'analyse statistique des données textuelles*, volume 2, pages 571–580. Presses Universitaires de Franche-Comté, 2006.

3.1 Introduction

According to Bauer (2001), an affix is productive if it is possible to create new words with it. An example of a productive affix in English is *-ness*. One can easily form new words ending in *-ness*. For instance, given the adjective *fractionated*, one can form the noun fractionatedness. By contrast, the suffix *-th*, as in *warmth*, is hardly productive, although an occasional neologism can be observed (Baayen, 2003).

There are several quantitative measures available for gauging the degree to which an affix is productive. An obvious measure is the size of the set of words containing the affix, henceforth its morphological category, as observed in a corpus. The more words an affix attaches to, the more productive that affix is. The disadvantage of this measure is that it does not take into account possible diachronic change in the productivity of an affix. So, a morphological category like that of the suffix *-ment*, which was more productive in the past, still has a considerable number of members, even though modern speakers are reluctant to use it in new words (Anshen and Aronoff, 1999). Conversely, speakers may also be reluctant to use an affix, even though it is fully productive in the sense that they could use it if required. An example is the Dutch suffix *-ster*, used to create nouns referring to female agents, such as *loop-ster*, 'female walker'. It is not fashionable in current Dutch to make the sex of the agent explicit, and the use of the unmarked counterpart with the suffix *-er* is preferred instead (Baayen, 1994b).

In order to overcome these difficulties, measures based on the Good-Turing estimate for unseen species (Good, 1953) have been introduced (Baayen, 1993). The measures that we will use here, first proposed in (Baayen, 1993), estimate the likelihood of observing a new formation with a given affix by counting the number of words that are observed only once, the hapax legomena, and calculating the proportion of such words with that affix. Since hapax legomena are relatively often new words and the number of new words created with a certain affix determines the productivity of that affix, hapax legomena are suited to predict the current productivity of affixes. Instead of measuring the extent to which a morphological category has been used in the past, this measure estimates the rate at which a morphological category is expanding and attracting new members.

Baayen and Renouf (1996) showed, on the basis of a large corpus study of British English, that hapax legomena are indeed the best estimators for the use of neologisms. Nishimoto (2003) compared productivity rankings obtained with the Good-Turing estimate with productivity rankings based on the deleted estimation method of Jelinek and Mercer (1985), and obtained similar rankings for both measures.

Most studies on productivity have proceeded on the implicit assumption that there would be an ideal speaker in a homogeneous speech community, whose knowledge is representative for all the other speakers in that commu-

nity. In the study of Bauer (2001), for instance, the possibility of variation in degrees of productivity across registers, social groups, and regions is not considered. Given the fact that such variation is involved in many linguistic variables (Biber, 1995; Keune et al., 2005), it is likewise expected to be involved in morphological productivity.

Previous variational studies of morphological productivity focused on how productivity varied with text type (Baayen, 1994a) and with register (Plag et al., 1999). Baayen (1994a) found that in some texts, like stories for children, the use of Germanic affixes is preferred, while in more official registers, Latinate affixes are most productive. Plag et al. (1999) showed that the productivity of a suffix may differ between written, formal spoken, and informal spoken language. Suffixes tended to be most productive in written language, and least productive in informal spoken language.

The aim of the present paper is to obtain further insight into the socio-geographic forces shaping morphological productivity in spoken Dutch. We investigated productivity as a function of whether a speaker lives in the Netherlands or in Flanders, of the speaker's sex, education level, and age.

3.2 Materials

We based our study on the Corpus of Spoken Dutch (CGN) (Oostdijk, 2002). This corpus consists of approximately 8.9 million words of spoken Dutch from various speech registers. These can be divided into three main registers, namely, spontaneous speech (unscripted conversations and telephone dialogues, 4.7 million words), speech from more formal settings such as debates, meetings, and interviews (3.3 million words), and read aloud speech of written Dutch (0.9 million words). As we were interested in exploring variation in spoken Dutch, we did not take into account the read aloud speech in the corpus. This left us with a corpus consisting of approximately 8.0 million words.

In the CGN, the characteristics of the speakers, for instance their home country, education level, sex, and age, are made available. This made it possible to address the socio-geographic variation in morphological productivity. To this end, we extracted 24 subcorpora according to a $2 \times 2 \times 2 \times 3$ factorial design with as predictors Country (the Netherlands versus Flanders, Education (High versus Non-High), Sex (Men versus Women), and Age (Young: < 40; Mid: 41–60; Old: > 60). The size of these subcorpora differed substantially, ranging from 27418 words (for old Flemish male speakers with a non high education level), to 942990 words (for middle aged Dutch male speakers with a high education level).

There are two slightly different criteria for what counts as a hapax legomena with a given affix. One criterion is to include only those hapax legomena with a given affix, for which that affix was attached to the word during the last morphological cycle. According to this criterion, the word *dank-baar-heid*

('grate-ful-ness') would be included in the count for *-heid* but not in the count for *-baar*. Another criterion is to include any word with that affix, including embedded words, as long as these embedded words are not present independently in the corpus either by itself or in other words. According to this criterion, *dank-baar-heid* would be included for the count of *-baar* if and only if *dank-baar* is not observed by itself. Gaeta and Ricca (2006) have shown that both criteria lead to very similar productivity rankings.

In order to facilitate extraction of the hapax legomena from the (only partially morphologically parsed) corpus, we selected those words in which affixes occurred either in the beginning or at the end of the word. We relaxed the first criterion by allowing words into our counts for which the affix is not attached during the last cycle, but only when they satisfied the second criterion. For instance, *pianospeler* ('piano player') fits our selection criteria given that *speler* is not present independently or as part of another complex word. The inclusion of these words did not substantially influence our results, but helped alleviate the problem of data sparseness.

The different affixes were selected on the basis of their existence in the morphologically parsed part of the CELEX lexical database (Baayen et al., 1993). Existing affixes were only used if there were ten of more word types in CELEX formed with that affix. Next, we restricted ourselves to the use of only those remaining affixes that appeared in the ANS grammar of Dutch (Geerts et al., 1984). In this way, 91 different affixes were selected for further analysis.

In order to determine the number of hapax legomena of these affixes as used in Dutch speech, we selected every word ending in the same characters as the affix from the CGN. So, for instance, for the prefix *be-* we selected al words starting in *be*. We designed a program that decided whether a word was a hapax legomena or was used more frequently, for instance in another inflectional form or as a part of another (often morphological complex) word. This program used the Memory-Based Morphological Analysis parser (Van den Bosch and Daelemans, 1999) that parses morphological complex words. The output of our program consisted of possible hapax legomena. We manually determined whether these words indeed contained the desired affix. For only 72 of the 91 selected affixes, hapax legomena occurred in the data set. In total 2251 hapax legomena were observed. The different affixes and their total number of hapax legomena are displayed in Table 3.1. In order to be able to measure the productivity of the affixes among the different subcorpora, we determined for each of the 2251 hapax legomena to which of the 24 subcorpora it belonged.

3.3 Method

The collected data posed a special challenge for statistical analysis for several reasons. First, many affixes emerged with zero counts for a large number of cells in the design. Second, each of the cells in the design contained counts based on subcorpora that differed in size by an order of magnitude. In order

Table 3.1: The 72 different affixes and their number of hapax legomena in The Corpus of Spoken Dutch

Affix	Frequency	Affix	Frequency	Affix	Frequency
aarts-	1	-erwijs	5	-ie	24
-ateur	1	-ief	5	-ist	24
hyper-	1	-es	6	-loos	24
-lijks	1	in-	6	ont-	25
opper-	1	-nis	6	-iteit	26
pseudo-	1	oer-	6	-aar	27
-uur	1	-zaam	6	-iseer	27
-dom	2	-erd	7	-atie	30
-ent	2	-matig	7	-erig	32
-erik	2	-air	9	-lijk	34
-st	2	-te	9	her-	37
tele-	2	-sel	10	-ij	41
-waarts	2	-ant	11	-baar	43
-aard	3	de-	11	be-	47
-elaar	3	-schap	12	super-	61
oud-	3	-aal	13	on-	64
psycho-	3	-eel	13	-isch	72
-weg	3	-ator	14	-achtig	90
-abel	4	inter-	14	-heid	100
bio-	4	-ling	14	ver-	114
-gewijs	4	-ster	20	-ing	141
-in	4	-isme	21	-er	175
co-	5	ge-	23	-ke	184
dis-	5	anti-	24	-je	477

to illustrate the diversity of the subcorpora, we displayed their size, the hapax frequency of their most productive affix, and the mean and the median of their total number hapax legomena in Table 3.2.

These size differences are due to the problems encountered by the builders of the Corpus of Spoken Dutch to obtain sufficient materials from non-highly educated speakers. Hence, any analysis based on the counts themselves, without taking the size of the subcorpora into account, would largely reflect the inequalities in the sizes of these subcorpora. Third, we needed to address the question whether to treat Affix as a fixed effect or a random effect. Since our sample is not exhaustive, one might argue that Affix is a random effect. On the other hand, we have sampled the most productive affixes, hence the sample is far from random, and might just as well be treated as fixed.

In the light of these challenges, we analyzed the data with three differ-

Table 3.2: The size of each subcorpus, the number of hapaxes of the most productive affix in the subcorpus, the mean and the median of the occurrences of the total number of hapax legomena in the subcorpus. Fl = Flanders, Nl = Dutch, H = High educated, NH = Non High educated, Y = aged < 41, M = aged 41–60, O = aged > 60

Subcorpus	Corpus Size	Max	Mean	Median
Nl male H Y	594692	47	2.4	1
Fl male H Y	450170	25	2.1	1
Nl male NH Y	234052	12	0.7	0
Fl male NH Y	122048	0	0.4	0
Nl female H Y	831388	71	2.8	0
Fl female H Y	554560	36	1.8	0
Nl female NH Y	318888	33	1.0	0
Fl female NH Y	128470	13	0.4	0
Nl male H M	942990	59	4.1	1
Fl male H M	574673	24	3.7	1
Nl male NH M	178167	11	0.6	0
Fl male NH M	52833	4	0.3	0
Nl female H M	481097	37	1.9	0
Fl female H M	424558	20	1.8	1
Nl female NH M	169749	14	0.5	0
Fl female NH M	51483	5	0.3	0
Nl male H O	344009	23	2.2	1
Fl male H O	283929	21	1.7	0
Nl male NH O	93095	3	0.2	0
Fl male NH O	27418	1	0.1	0
Nl female H O	166320	17	0.8	0
Fl female H O	132367	16	0.7	0
Nl female NH O	182288	22	0.6	0
Fl female NH O	38865	4	0.2	0
Total	7378109	71	1.3	0

ent statistical techniques. Our first model was obtained using ordinary least squares regression with the proportion of hapax legomena in the subcorpus as dependent variable. The statistic formula is given below:

$$E[Y] = X\beta + \varepsilon,$$

in which Y represents the criterion, X the predictors and β the weights.

We rescaled these proportions by multiplying them by 100000, and raised them to the power of 0.25 in order to reduce the skew in their distribution. Since proportions for large subcorpora are more reliable than proportions for small subcorpora, we fitted the ordinary least squares model to the data using the sizes of the subcorpora as weights. In this model, we treated Affix as a fixed effect. The results for a model containing only simple effects are shown in the left section of Table 3.3, the results for a model in which two-way interactions were allowed are listed in Table 3.4.

We also analyzed the data with a linear mixed effects model with Affix as random effect, using the same transformed proportions as in the ordinary least squares regression. The statistic formula looks as follows:

$$E[Y] = X\beta + Zb + \varepsilon,$$

in which Y represents the criterion, X the data matrix, β the coefficients of the fixed effects, Z a copy of the data matrix, and b the coefficients of the random effects.

We used the *lme4* library of Bates and Sarkar (2005), using restricted maximum likelihood estimation. The *lme4* library provides improved algorithms compared to the *nlme* library of Pinheiro and Bates (2000), but has the disadvantage that it is still under development. At the time of writing, it was not possible for us to make use of weighted models. The results obtained with this multilevel model are listed in the central sections of Tables 3.3 and 3.4.

Our third model made use of a generalized linear model with a binomial link function, of which the statistic formula is:

$$E[Y] = 1/(1 + e^{-X\beta}),$$

in which Y represents the criterion, X the predictors and β the weights.

Hapax legomena were considered as successes, and all remaining words in the subcorpus were counted as failures. In this logistic model, the total number of words in the subcorpora is automatically included as weight. The third sections of Tables 3.3 and 3.4 summarize the results obtained. As we were coping with already fairly sparse data, we did not take three-way interactions into account in any of these analyses.

Table 3.3: F and p statistics for three simple main effects models: an ordinary least squares model (lm), a multi level model (lmer), and a generalized linear model (glm). For lm, $df = 1651$, for lmer, $df_2 = 1722$

	lm			lmer			glm			
	F	df_1	p	F	df_1	p	F	df_1	df_2	p
Country	8.07	1	0.0046	0.02	1	0.8898	18.94	1	1726	< 0.0001
Education	182.29	1	< 0.0001	207.02	1	< 0.0001	22.10	1	1725	< 0.0001
Sex	58.93	1	< 0.0001	21.57	1	< 0.0001	34.33	1	1724	< 0.0001
Age	13.70	2	< 0.0001	12.04	2	< 0.0001	16.79	2	1722	< 0.0001
Affix	27.87	71	< 0.0001				64.02	72	1651	< 0.0001
	$R^2 = 0.42$			$R^2 = 0.45$			$R^2 = 0.73$			

3.4 Results

All three models revealed highly significant simple main effects for Education Level, Sex, Age, and Affix. In the ordinary least squares model and in the logistic regression the effect for Country was also significant. Highly educated older men revealed the greatest overall productivity. As expected, productivity varied substantially from affix to affix. The prefix *pseudo-* turned out to be least productive, and the diminutive suffix *-je* to be most productive. For each of the three models, we calculated the squared correlation of the observed and expected cell counts. The resulting R^2 was largest for the logistic regression model (0.73), and substantially smaller for the other two models (0.42 and 0.45).

When we allowed two-way interactions into the ordinary least squares model (see Table 3.4), many interactions emerged as significant, and the R^2 increased from 0.42 to 0.73. For the multilevel model, the addition of two-way interactions led to only a small improvement in the R^2 from 0.45 to 0.49. The generalized linear model with two-way interactions emerged as most successful, with an increase in the R^2 from 0.73 to 0.95.

The substantially better fit achieved with the generalized linear model is due to two factors. First, inspection of the residuals shows that the generalized linear model is more successful in predicting the zero counts. The generalized linear model is not constrained by the normality assumption that governs the distribution of the residuals in ordinary least squares regression. Second, the disappointing performance of the linear mixed effect model is due to the Zipfian nature of affix productivity. Linear mixed effect models assume that random effects follow a normal distribution with mean zero and unknown variance. When we include Affix as a random effect in the multilevel model, we implicitly assume that the difference in productivity of a given affix compared to the average productivity of an affix is normally distributed. This distribution, however, is decidedly non-normal. This explains the disappointing performance of the linear mixed effect model: it is simply not appropriate for our kind of data. For the discussion of the interactions, we therefore restrict ourselves to the logistic regression model.

Table 3.4: F and p statistics for three models allowing two-way interactions: an ordinary least squares model (lm), a multi level model (lmer), and a generalized linear model (glm). For lm, $df_2 = 1290$, for lmer, $df_2 = 1722$

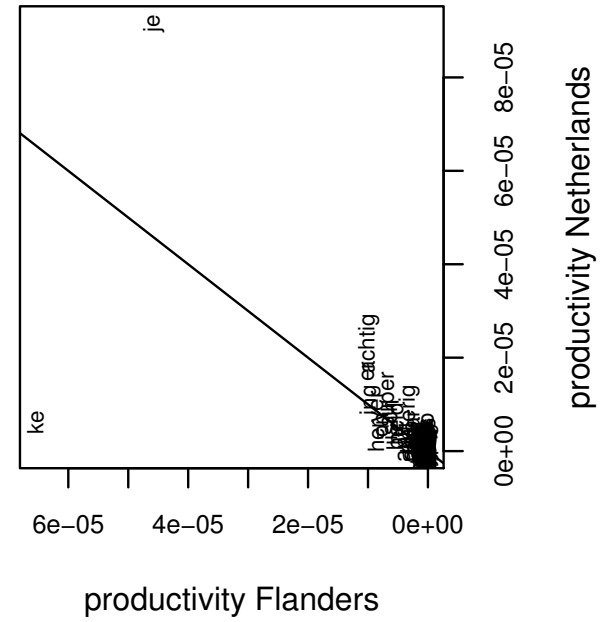
	lm			lmer			glm			
	F	df_1	p	F	df_1	p	F	df_1	df_2	p
Country	10.53	1	0.0012			n.s.	18.94	1	1726	< 0.0001
Education	237.78	1	< 0.0001	118.28	1	< 0.0001	22.10	1	1725	< 0.0001
Sex	76.86	1	< 0.0001	25.22	2	< 0.0001	34.33	1	1724	< 0.0001
Age	17.88	2	< 0.0001	10.54	1	< 0.0001	16.79	2	1722	< 0.0001
Affix	36.35	71	< 0.0001			n.s.	64.02	71	1651	< 0.0001
Country:Sex	4.65	1	0.0313			n.s.	5.69	1	1293	0.0170
Country:Age	9.75	2	< 0.0001			n.s.	4.95	2	1294	0.0070
Educ:Sex	16.70	1	< 0.0001	28.27	1	< 0.0001				n.s.
Sex:Age	4.11	71	0.0167			n.s.				n.s.
	F	df_1	p	F	df_1	p	F	df_1	df_2	p
Country:Affix	3.54	71	< 0.0001	38.42	2	< 0.0001	5.37	71	1580	< 0.0001
Educ:Affix	2.01	71	< 0.0001	26.97	3	< 0.0001	2.10	71	1509	< 0.0001
Sex:Affix	2.32	71	< 0.0001			n.s.	2.43	71	1438	< 0.0001
Age:Affix	1.80	142	< 0.0001	15.65	4	0.0079	1.80	142	1296	< 0.0001
	$R^2 = 0.73$			$R^2 = 0.49$			$R^2 = 0.95$			

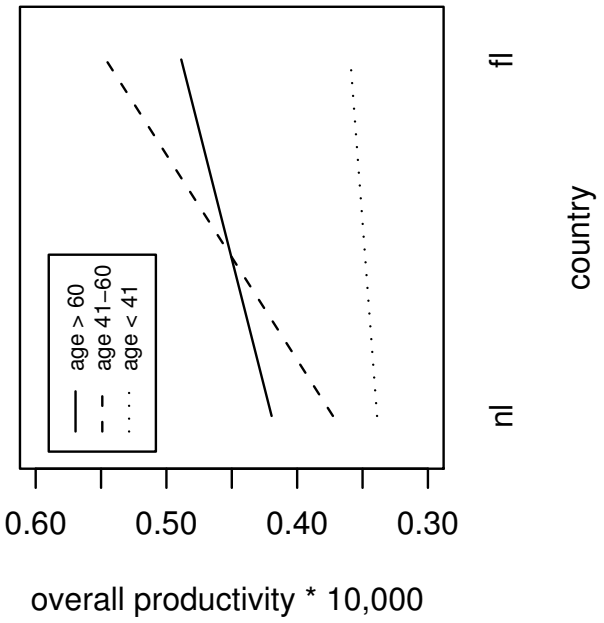
The interaction of Country by Sex ($F(1, 1293) = 5.69$, $p < 0.0170$) indicates that in both the Netherlands and Flanders women use affixes less productively then men. The interaction of Country by Age ($F(1, 572) = 10.31$, $p < 0.0013$) is illustrated in the upper left panel of Figure 3.1. Affixes were used less productively by speakers aged between 19 and 40 than by speakers above 40 ($F(1, 1723) = 28.52$, $p < 0.0001$). The interaction of the subset of speakers with age above 40 was also significant ($F(1, 1790) = 6.89$, $p < 0.0087$). While in the Netherlands speakers above 60 use affixes more productively, in Flanders they use them less productively compared to middle aged speakers. In other words, in the Netherlands productivity increases with age, while in Flanders, the old age group is intermediate between the young and middle age group. The relatively low productivity for older speakers in Flanders may be due to the fact that Dutch was not the official language in Flanders until 1963 (Geeraerts et al., 1999). For these speakers, Dutch is somewhat more like an official register in which they are less fluent, and less productive. However old speakers from Flanders use affixes less productively then middle-aged speakers, they still use them more productively then Dutch speakers. This is probably due to the fact that Flemish speakers have an additional vocabulary (Southern Dutch, Flemish), while Dutch speakers only use the standard (Northern) vocabulary (e.g., Geeraerts et al., 1999).

The productivity of the affixes also varied from affix to affix for all four predictors, as witnessed by the interactions of Country by Affix ($F(71, 1580) = 5.37$, $p < 0.0001$), Education by Affix ($F(71, 1509) = 2.10$, $p < 0.0001$), Sex by Affix ($F(71, 1438) = 2.43$, $p < 0.0001$), and Age by Affix ($F(142, 1296) = 1.80$, $p < 0.0001$).

We visualized the interaction of Country by Affix in the upper left and lower left and right panels of Figure 3.1. Thanks to the use of contrast coding, with contrasts being made between a given affix and the least productive affix (which was *pseudo-*), the coefficients of Affix and of Affix by Country provide a straightforward estimate of differences in degrees of productivity within and across two countries. The plots are calibrated for young highly educated women.

In the upper right panel the two most productive affixes, the diminutives *-je* and *-ke* are clearly differentiated: *-ke* appears in the upper left corner, which means that it is more productive in Flanders, while *-je* appears in the upper right corner, indicating that it is used more productively in the Netherlands. This is exactly as expected, as these two forms of the diminutive are well-known regional markers (Geerts et al., 1984).


je
ke
6e−05
4e−05
2e−05
0e+00
0e+00
2e−05
4e−05
6e−05
8e−05
productivity Flanders
productivity Netherlands


age > 60
age 41−60
age < 41
0.60
0.50
0.40
0.30
overall productivity * 10,000
nl
fl
country

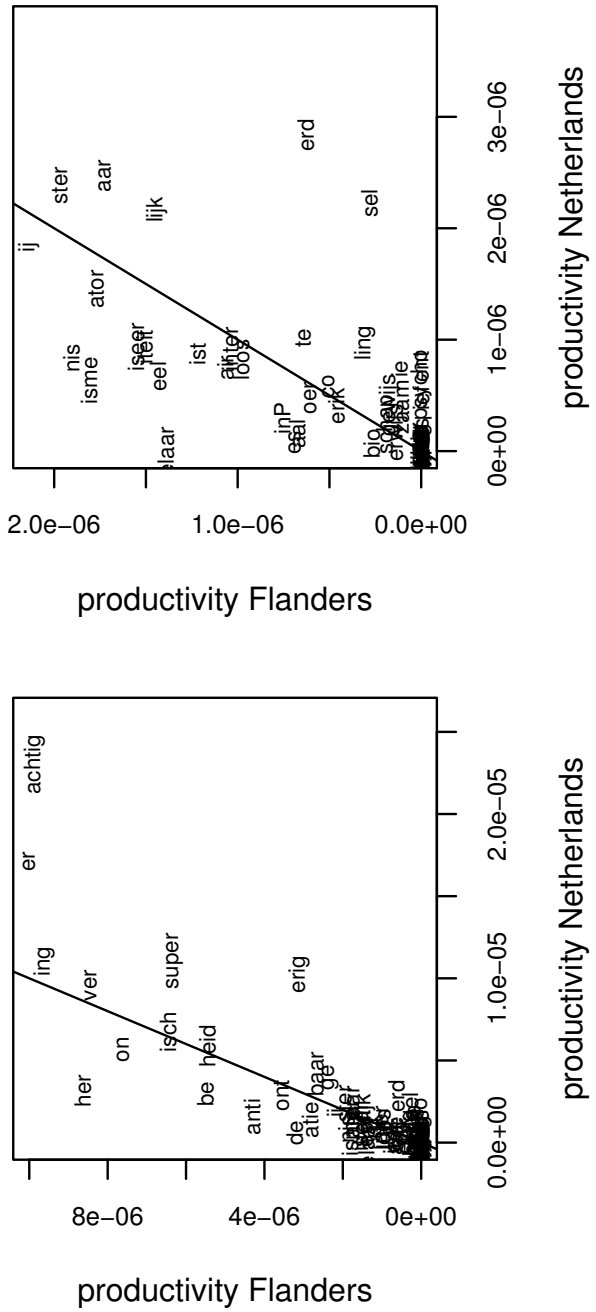


Figure 3.1: The upper left panel illustrates the interaction of country by age. The y-axis depicts the overall productivity of the affixes times 100000. The remaining panels visualize the by-affix adjustments for Country. The x-axes represent the productivity of the affixes for young women from the Netherlands, the y-axes their productivity for young women from Flanders. The lower panels are close ups of the upper right panel

The diminutives enjoy the greatest productivity in spoken Dutch of all our affixes. In order to visualize the structure of the cluster in the lower left hand corner, we zoomed in on this part of the plot, resulting in the lower left panel. This panel reveals that the suffixes *-erig*, *-er*, and *–achtig,* and the prefix *super-* are more productive in the Netherlands, while the prefixes *her-*, *anti-*, *be-*, and *on-* are more productive in Flanders. The lower right panel zooms in on the cluster of least productive affixes. The suffix *-atie* was reported by Pauwels (1964) to be more productive in Flanders, and his conclusion is supported by our data: *-atie* is located above the $Y = X$ line.

In summary, our multivariate approach to variation in morphological productivity succeeds not only in capturing regional differences already known from the previous literature to exist (*-je* versus *-ke*, *-ing* versus *-atie*), but also offers the possibility to explore many potential carriers of socio-geographic variation simultaneously.

3.5 Conclusions

We have shown that it is possible to chart variation in morphological productivity across socio-geographic dimensions, even when there are substantial differences in the sample sizes underlying the counts in the cells of the statistical design. We obtained excellent results with a generalized linear model with a binomial link, even though the success probabilities in our data were extremely small. Given the possibilities for visualization of the variation in the use of the individual affixes, we believe the present approach offers a useful alternative to correspondence analysis for count data in cells with different underlying sample sizes.

References

Anshen, F. and M. Aronoff, 1999. Using dictionaries to study the mental lexicon. *Brain and Language*, 68: 16–26

Baayen, R. H., 1993. On frequency, transparency, and productivity. In G. E. Booij and J. van Marle, eds., *Yearbook of Morphology 1992*. Kluwer Academic Publishers, Dordrecht, 181–208

Baayen, R. H., 1994a. Derivational productivity and text typology. *Journal of Quantitative Linguistics*, 1: 16–34

Baayen, R. H., 1994b. Productivity in language production. *Language and Cognitive Processes*, 9: 447–469

Baayen, R. H., 2003. Probabilistic approaches to morphology. In R. Bod, J. Hay and S. Jannedy, eds., *Probabilistic linguistics*. The MIT Press, 229–287

Baayen, R. H. and A. Renouf, 1996. Chronicling The Times: Productive Lexical Innovations in an English Newspaper. *Language*, 72: 69–96

Bates, D. M. and D. Sarkar, 2005. The lme4 library. *[On-line], Available: http://lib.stat.cmu.edu/R/CRAN/*

Bauer, L., 2001. *Morphological productivity.* Cambridge University Press, Cambridge

Biber, D., 1995. *Dimensions of Register Variation.* Cambridge University Press, Cambridge

Gaeta, L. and D. Ricca, 2006. Productivity in Italian word formation: a variable-corpus approach. *Linguistics*, 44: 57–89

Geeraerts, D., S. Grondelaers and D. Speelman, 1999. *Convergentie en Divergentie in de Nederlandse Woordenschat. Een Onderzoek naar Kleding- en Voetbaltermen.* Meertens Instituut, Amsterdam

Geerts, G., W. Haeseryn, J. de Rooij and M. C. van den Toorn, 1984. *Algemene Nederlandse Spraakkunst.* Wolters-Noordhoff, Groningen

Good, I. J., 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40: 237–264

Jelinek, F. and R. Mercer, 1985. Probability distribution estimation for sparse data. *IBM technical disclosure bulletin*, 28: 2591–2594

Keune, K., M. Ernestus, R. van Hout and R. H. Baayen, 2005. Social, geographical, and register variation in Dutch: From written 'mogelijk' to spoken 'mok'. *Corpus Linguistics and Linguistic Theory*, 1: 183–223

Nishimoto, E., 2003. Measuring and computing the productivity of Mandarin Chinese suffixes. *Computational Linguistics and Chinese Language Processing*, 8 (1): 49–76

Oostdijk, N. H. J., 2002. The Design of the Spoken Dutch Corpus. In P. Peters, P. Collins and A. Smith, eds., *New Frontiers of Corpus Research.* Rodopi, Amsterdam, 105–112

Pauwels, J. L., 1964. Woorden op -atie en -ering in het nederlands. *Verslagen en Mededelingen van de Koninklijke Vlaamse Academie voor Taal- en Letterkunde*: 205–210

Pinheiro, J. C. and D. M. Bates, 2000. *Mixed-effects models in S and S-PLUS.* Statistics and Computing, Springer, New York

Plag, I., C. Dalton-Puffer and R. H. Baayen, 1999. Morphological productivity across speech and writing. *English Language and Linguistics*, 3 (2): 209–228

Van den Bosch, A. and W. Daelemans, 1999. Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99*. University of Maryland, USA, 285–292

CHAPTER 4

Derivational and lexical productivity across written and spoken Dutch[1]

Abstract

In this study we explored variation patterns in derivational and lexical productivity in both written and spoken Dutch. The explanatory variables are language *use*, i.e. register, and characteristics of the language *user*, i.e. country, gender, education level, and age. It turned out that both *use* and *user* related factors have a significant impact on the productivity found in texts. Variation patterns in derivational productivity were mirrored by variation patterns in lexical productivity. There is however no derivational pattern pertaining to all affixes. Variation patterns are the outcome of affix-specific and even word-specific frequency patterns.

Keywords: lexical productivity, derivation, affix, register variation, sociolinguistic variation, corpus analysis, written and spoken text analysis

4.1 Introduction

Lexicons differ in their productivity to coin new words. Huge corpora are available nowadays that represent different registers, genres, styles, and groups of speakers. These corpora can be analyzed to determine the potential productivity of the lexicons on which they are rooted. Their potential productivity can be measured by counting the number of hapax legomena, i.e. the words only occurring once, divided by the total number of tokens in the corpus (cf. Baayen, 2009). The outcome estimates the growth rate of the lexicon in question, i.e. how many words can be expected to show up that have not been observed

[1] This study, co-authored by Roeland van Hout and Harald Baayen, has been submitted for publication.

before when the corpus expands in size. A high growth rate is indicative of a lexicon that still contains or may produce many words that were hitherto not observed, a zero growth rate indicates that no unattested words were left, as all words were being used. The growth rate will decrease the more the sample of tokens (the corpus) grows.

Productivity can be established for a lexicon as a whole, making no distinctions between different morphological categories and semantic fields. In this case, any unattested word is a candidate to become a new hapax legomenon. The subclass of words that is probably most relevant to investigate in relation to productivity, next to compounds, is that of derivational forms, as affixes may create new words. Plag, Dalton-Puffer and Baayen (1999b) analyzed the occurrence of 15 English derivational suffixes, which were classified as at least moderately productive, in the British National Corpus. Not only the productivity of the different suffixes varied enormously, the suffixes also showed a non-uniform productivity pattern across registers: a suffix can be highly productive in one register, and hardly productive in another, while another suffix shows the opposite pattern.

Another essential observation is that the productivity of affixes often changes over time. An affix has by definition a stage where it is somehow productive as a word formation device. The degree of productivity may become less over time and even stop, leaving a fixed set of words in the lexicon marked by that affix. The English affixes *of* and *at*, for instance, ceased to be productive (Anshen and Aronoff, 1997). A good example of decreasing productivity in Dutch is the derivational suffix *-lijk*, as in *natuurlijk* ('nature like', nowadays most frequently meaning 'of course') (Keune et al., 2005). On the other hand, new affixes may show up. For example, the noun *gate* became a suffix, in nouns denoting an actual or alleged scandal, following the Watergate scandal in 1972. Some affixes became (more) productive over the last decennia, for instance the prefixes *super, mega* and *giga*, both in English and Dutch.

Affixes may vary in their degree of productivity, but given their potential to create new words, the class of derivations as a whole seems to have a privileged status in producing hapax legomena. Several studies investigated the particular role of derivations. Baayen and Renouf (1996) explored the degree of productivity of five derivational affixes in a British newspaper corpus. Baayen and Neijt (1997) studied the productivity of the Dutch suffix *-heid* (equivalent to the English suffix *-ness*) in a newspaper corpus and found that it was most productive in the sections on literature and art, and not productive in articles on economics. Baayen (1994) showed that derivational affixes were used to a different extent in spoken and written language, with an obviously higher frequency in written language. As mentioned before, Plag, Dalton-Puffer and Baayen (1999a) investigated variation across speech and writing. The overalll affix productivity was highest in written and lowest in informally spoken language. Biber (1988) and Biber and Conrad (2009) investigated 67 linguistic features to investigate register variation in English with the help of 67 typical linguis-

tic features. There was one morphological feature, namely 'nominalizations', comprising the derivations *-tion, -ment, -ness, -ity* (Biber, 1988: 227). These derivations loaded on one of the dimensions distinguished by Biber, namely that of 'Situated versus Elaborated Reference' (Biber, 1995: 155), where they loaded high on the Elaborated Reference side, indicating that they occur more in abstract registers.

What kind of results were found for global lexical productivity? This measure was studied in a number of studies as well. Smith and Kelly (2002) used lexical productivity in a stylometric study to distinguish author characteristics. Van Gijsel et al. (2005, 2006) and Van Gijsel (2007) showed that register was the principal factor in explaining differences in lexical productivity in spoken Dutch, country (the Netherlands vs. Flanders (Belgium)), gender and age were also relevant but less pervasive factors. Keune et al. (submitted) used lexical productivity to investigate variation among groups of speakers of Dutch. They distinguished speakers by country, and the social factors gender, education level, and age, and demonstrated the importance of investigating country specific and social patterns of lexical productivity in samples of spoken Dutch. They found, for instance, that in spontaneous speech men revealed a higher degree of lexical productivity than women, and that in Flanders highly educated speakers were lexically more productive than lower educated speakers. Härnqvist et al. (2003) found differences in lexical productivity for gender and education level in a corpus containing 415 interviews with Swedish men and women. Speech of men and highly educated speakers appeared to be most productive.

What could be the relationship between lexical and derivational productivity? Is it a stable relationship or is it influenced by sociolinguistic factors as they have a different impact on word formation processes, and, concerning derivational productivity, is there one overall sociolinguistic variation pattern for all affixes or are there affix-specific patterns? Whereas a lexicon seems to be defined better as a relatively closed set of lexical elements, derivations seem to be marked more by their open character as they have the potential to produce new words. However, affixes may cease to be productive, implying that derivations become unanalyzed lexical elements of the lexicon, having the same status as all other lexical elements (Keune et al., 2005). Moreover, the formation of new words is not the exclusive domain of derivations. Productivity comprises a range of different word formation processes and is enacted and enforced by all these processes. The use of affixes obeys normal lexical selection rules and the derivational productivity of non-productive affixes (in a way a contradictio in terminis) mirrors the rules of common lexical productivity. To complicate matters even more, a lexicon has other means than derivations to produce new words (neologisms). New word forms are not only created by the use of derivational affixes, but also by word compounding and word composition. Other productive processes that makes new words enter the lexicon are word borrowing (cf. Chesley and Baayen (2011) for a study on entrenchment into the

lexicon of lexical borrowings) and word blending (cf. Gries (2003) for a discussion of blends in English). As a consequence of all these processes that allow new words to enter the lexicon, each lexicon is somehow productive. This is reflected in higher lexical productivity scores, but it may be mirrored at the same time by higher derivational productivity scores. That would link up lexical and derivational productivity.

One may ask the question what could connect or chain all affixes of a language, as the set of affixes ranges over the full scale of non-productivity towards strong productivity. One may argue that affixes do not constitute a coherent, structured set, but that they are the accidental outcome of various innovative lexical processes, accidently sharing the property of having a derivational format. The format may be language specific, in the sense that some languages prefer prefixes, whereas other languages prefer suffixes or some mixture, but that relates to preferences in form. The full set, however, does not seem to be bound by semantic or more general principles. Each derivation is having its own, autonomous lexical distribution and degree of productivity. The complete set is heterogeneous, part of the affixes perhaps fitting the overall pattern of lexical productivity, especially those that are no longer productive or are only very moderately productive. Productive affixes will be a substantially different set, having again their own distribution pattern, and their own rate of productivity. This would mean that the claim that high derivational productivity is more typical of written speech has no theoretical motivation. On the one hand this may depend on the affixes investigated, and on the other hand on the influence of sociolinguistic variables.

The aim of the present study is to reveal variation patterns determined by language *use*, i.e. register, and by the language *user* i.e. the factor of country and social factors in derivational and lexical productivity in both written and spoken Dutch, and to investigate whether these two forms of productivity are similar or different. We will investigate derivational productivity as a whole, including all word forms that can be classified as derivations. The individual Dutch affixes constitute a separate factor in our data analyses, and we want to know whether differences are affix-specific or that we may generalize to subsets of affixes, or, perhaps, to the whole set. Our prediction is that both forms of productivity are highly comparable in their capacity to produce hapax legomena. We predict high correlations within registers between lexical and derivational productivity.

For written Dutch, we explored variation according to the *use* of the language, by comparing three different newspaper registers, namely quality newspapers, aiming at readers with a higher education level, tabloids (or national newspapers), and regional newspapers, and according to *user* by comparing two newspaper countries, namely the Netherlands and Flanders. For spoken Dutch we also compared variation according to the *use* of the language by comparing different speech registers: formal monologues, formal dialogues, and informal dialogues. We explore variation according to the *user* not only by country (the Netherlands versus Flanders (the Dutch speaking area of Belgium)), but also

by the social factors of gender, education level, and age. This was possible due to the availability of speaker characteristics in the Spoken Dutch Corpus.

We start the analysis of the corpora data with comparing derivational and lexical productivity in the main subcorpora of the written and spoken Dutch corpora, in order to give a first answer to the question whether derivational productivity generalizes to lexical productivity. We investigate whether the resemblance of the derivational productivity between written and spoken Dutch is an overall effect or whether there are affix-specific differences. Next, we include the country factor and, for spoken Dutch, social factors in our analyses to investigate the influence of sociolinguistic factors on the distribution of hapax legomena in the corpora.

4.2 Method

4.2.1 Written Dutch

We used the CONDIV corpus (Grondelaers et al., 2000) to investigate productivity in written Dutch. This corpus contains written Dutch taken from six different newspapers. Three of these newspapers are published in Flanders (*De Standaard*, *Het Laatste Nieuws*, and *Het Belang van Limburg*), and three in the Netherlands (*NRC Handelsblad*, *De Telegraaf* and *De Limburger*). *De Standaard* and *NRC Handelsblad* are quality newspapers aiming at readers with a higher level of education. *Het Laatste Nieuws* and *De Telegraaf* are national tabloids with a populistic editorial policy, aiming at a broad readership. *Het Belang van Limburg*, and *De Limburger* are regional newspapers publishing a mixture of (inter)national and regional news. They aim at a general readership in the Limburg region of respectively Flanders and the Netherlands. This set of six newspapers gives a 2 by 3 orthogonal design, cross-classifying country (Flanders versus the Netherlands) and register (quality, tabloid, regional).

The smallest newspaper subcorpus comprised approximately 1.5 million words. We selected samples with the same size from the other newspapers by taking the first 1.5 million words, the result being six subcorpora of 1.5 million words and a full or total corpus of 9 million words.

We started with a selection of 91 affixes to investigate derivational productivity. This selection was based on two criteria. First, the affix was listed as a constituent for at least ten different word types in the morphologically parsed section of the CELEX lexical database (Baayen et al., 1995), to exclude affixes that became obsolete. Second, the affix was listed in the ANS reference grammar of Dutch (Geerts et al., 1984). Two variants of the diminutive affix were added to the affix list, as they turned out to be used frequently in our corpus. The diminutive suffix of standard written Dutch *-je* has an additional Belgian Dutch variant *-ke* and an additional Netherlandic Dutch variant *-ie*.

For each affix, we counted the number of words with that affix that occurred once only in the full corpus of the six newspapers. These counts of the deriva-

tional hapax legomena were obtained as follows. We first extracted all words potentially beginning with one of the prefixes or ending in one of the suffixes in our set of selected affixes by means of a string search. In particular for shorter affixes such as *-ie* and *be-*, the string search returned many spurious, pseudo-affixed words such as *familie* ('family') in the case of *-ie* and *beest* (beast) in the case of *be-*. We used a memory-based morphological parser (Van den Bosch and Daelemans, 1999) to remove most pseudo-affixed words from our list. Next, we manually removed the remaining pseudo-affixed candidates.[2] A word was accepted as a hapax legomenon if it occurred as a separate word, not being part of another complex word, or if it only occurred in a complex word that itself was a hapax legomenon. For three affixes, no hapax legomena were found. These affixes were removed from the data set. This means that our analyses of written Dutch are based on a set of 88 affixes occurring in 2,403 hapax legomena. For each hapax legomenon, we registered its occurrence in the six subcorpora, the resulting data matrix containing 528 cells (six newspapers by 88 affixes).

The lexical productivity was investigated on the basis of the occurrence of lexical hapax legomena. We first extracted the hapax legomena for the joint corpus of 9 million words. Filters were used to remove hapax legomena that were only mark-up language and words containing numbers. The presence of many typos led us to manually inspect 1,000 randomly selected hapaxes for each newspaper, in order to estimate the corresponding proportion of typos. With these estimates, we corrected our raw counts. This resulted in 115,483 hapax counts, ranging from 15,599 to 22,345 hapax legomena in the six newspapers. These counts exceed the counts of derivational hapax legomena by at least one order of magnitude (the total number of derived hapax legomena for all newspapers was 2,403). In other words, the contribution of derivational hapax legomena to the overall count of hapax legomena is negligible (about 2%). This allows us to view the overall counts of the lexical and the derivational hapax legomena as separate measures of productivity.

4.2.2 Spoken Dutch

In order to investigate derivational and lexical productivity of spoken Dutch, we made use of the Spoken Dutch Corpus (CGN) (Oostdijk, 2002). This corpus consists of approximately 8.9 million words of spoken Dutch from various speech registers. The spontaneous part of the corpus (8.0 million words) can be divided into private speech (unscripted conversations and telephone dialogues: 4.7 million words) and public speech (3.4 million words). The public speech data can be split up in dialogues (for instance debates, meetings, and interviews: 2.3 million words), and monologues (news, reportages, and commentaries (all broadcast), reviews, ceremonious speeches, and lectures: 1.1 million words).

[2]We checked whether our procedure (parser, followed by manual selection) was correct by manually selecting all hapaxes for 10 of the 91 affixes. The results were the same for both procedures.

The remaining, non-spontaneous part of the CGN comprises read aloud speech of written Dutch from the library of the blind (0.9 million words). As our aim was to explore variation in spoken Dutch, we excluded this last speech register, which left us with a corpus of 8.0 million words. The Spoken Dutch Corpus comes with detailed meta-data about the speakers, from which we extracted the speaker's country (the Netherlands versus Flanders), and the speaker's social factors of gender, education level (high versus non high), and age (Young: < 40; Mid:$41 - 60$; Old: > 60). On the basis of these factorial contrasts plus the contrast between private speech, public dialogues, and public monologues we have a $2 \times 2 \times 2 \times 3 \times 3$ factorial design, containing 72 subcorpora.

The metadata turned out to be unknown for part of the public speech data. This data was excluded. We therefore worked with a total corpus size of 7.4 million words consisting of 4.7 million words from private dialogues, 1.9 million words from public dialogues, and 0.8 million words from public monologues. The sizes of these 72 subcorpora differed substantially. For three subcorpora there was no data available (all three public monologues from lower educated speakers). The largest subcorpus, containing private speech from highly educated young Dutch women, consisted of 727,102 words.

We studied the same 91 affixes as in our analyses of written Dutch, and counted the number of words with that affix occurring once only in the full corpus of 7.4 million words, using the same selection criteria. For 19 affixes, we did not observe any hapax legomena. As a consequence, our analyses are based on the data of 73 affixes. These 73 affixes occurred in 2,325 hapax legomena, a number comparable to the hapax legomena found in the written corpus (2,403 hapax legomena). We cross-tabulated these hapax legomena by subcorpus and affix, resulting in a data matrix with 5,256 cells (72 subcorpora times 73 affixes). For each cell count, we registered the size of the subcorpus and its associated variables (country, gender, education level, age, speech register, and affix).

To investigate lexical productivity, all hapax legomena were extracted from the full corpus of 7.4 million words, and they were classified for the subcorpus they belonged to. The full corpus comprised 50,532 hapax legomena counts, a number that is clearly lower than in the written corpus (with number of hapax legomena of 115,483). The hapax counts ranged from 0 for the subcorpora containing between 0 and 143 words to 3,624 for the largest subcorpus containing 727,102 words. The contribution of the derivational hapax legomena to the overalll count of hapax legomena is about 4%, which is higher than the 2% of the written corpus, though low enough to consider the overalll counts of the lexical and the derivational hapax legomena as separate measures of productivity.

Because of the small ratio of hapax legomena occurring in the corpora, and the substantial size differences of the subcorpora, we report the number of hapax legomena as parts per million (ppm): the number of hapaxes expected in a (sub)corpus of 1 million words. The ppm supports comparing in a straightforward way differently sized (sub)corpora.

4.3 Results

We first give the overall results of the derivational and lexical productivity for the three main divisions in the written (three newspaper categories) and spoken (dialogue - private speech, dialogue - public speech, monologue - public speech) corpus. The results are given in Figure 4.1, the y axis representing the derivational productivity, the x axis the lexical productivity, measured in the number of hapaxes per million words.

Figure 4.1 shows a linear pattern within each of the two corpora, in which, as mentioned before, the size of the lexical productivity exceeds the size of the derivational productivity by far. Within the spoken corpus, the lexical hapaxes seem to increase regularly in relation to the derivational hapaxes, giving an expected order of productivity running from lower scores for dialogues in private speech towards higher scores for monologues in public speech. The more formal and prepared public monologues are by far more the most productive category in the spoken corpus.

The three news paper categories show a linear increase as well, in the expected order, going from the tabloids to quality newspapers, the latter ones having higher outcomes for lexical and derivational productivity. The newspapers have higher scores for the lexical hapaxes than the spoken corpora, except for monologues. The high score of the monologues may be explained by the relatively small size of this subcorpus in combination with its high resemblance to written Dutch, as it is prepared speech. It is clearly different from the dialogues. Another conclusion to be drawn from Figure 4.1 is that the number of derivational hapaxes in speech are not lower than in written language. This may be the consequence of having selected the full set of Dutch derivations, and not a particular subset as was done in earlier research (e.g. Plag, Dalton-Puffer and Baayen, 1999a). Scatter plots of the separate affixes can reveal further details about the contribution of separate affixes and their distribution.

Figure 4.2 gives an overview of the affix-specific productivity in written compared to spoken Dutch in ppm. The left panel exemplifies that the diminutive suffix *-je* enjoys the highest degree of productivity in both written and spoken Dutch. Its location is however far to the left of the line $Y = X$ (represented by the solid line) in the spoken corpus. The diminutive affix occurs more frequently in spoken Dutch. The same can be said about its Belgian Dutch variant *ke*. The suffix *-ing* (*handeling*, 'action'), by contrast, is more productive in written Dutch.

The right panel zooms in on the cluster of affixes in the lower left corner of the left panel, which entails the less productive affixes. Affixes typically used in written Dutch turn out to be *-ster*, *oud-*, and *-aar*, but four other affixes *ver-*, *-achtig*, *on-*, and *super-* are more typical for spoken Dutch. The Netherlandic Dutch variant of the diminutive, *-ie*, can be found more down in the left corner. This variant of the diminutive is more typical for spoken Dutch as well. Another example of an affix typically for spoken Dutch is the affix *super-*. This is an

intensifier, a word type that can be typical for spoken language (cf. Tagliamonte 2008).

We decided to exclude the diminutives from our quantitative analyses, as the diminutives were most often a deviant category in the set of affixes, as they were highly influential within the category of affixes in analyzing hapaxes because of their numerosity, which indicates their productivity. Including their country specific variants, it was the only set of affixes with such a high impact being caused by effects for country, that were not present for the remaining group of affixes, and had social effects. The domain of application of the Dutch diminutives is larger than the set of nouns, as it may apply to adjectives, adverbs and prepositional phrases as well. Its vast domain of application makes this affix radically different from any other affix in Dutch, with an unmatched productivity. It is the only derivational category that is studied systematically in first language acquisition, emphasizing its special status, not only in Dutch. Savieckienė and Dressler (2007) edited a book that gives a cross-linguistic perspective, including Dutch, on the acquisition of diminutives. They conclude that diminutive formation often tends to be the first pattern of word formation to emerge, with a high degree of productivity and transparency. Its special productive status is mirrored in the presence of its two country variants *-ke* and *-ie* in our data set. None of the other affixes had country variants. We kept the diminutives in the plots (cf. Figures 4.4 and 4.6) on the individual affixes to provide additional information on the distributional properties of this suffix.

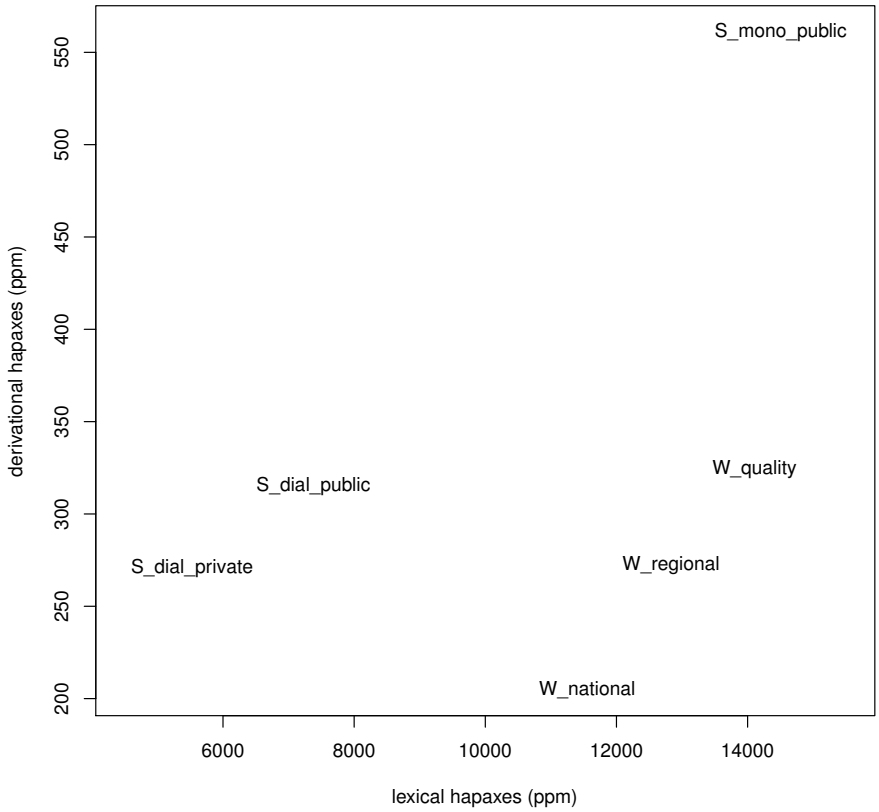


Figure 4.1: Scatterplot of the lexical versus the derivational complexity in the main subcorpora of written (W) and spoken (S) Dutch. The main divisions are dialogue - private speech, dialogue - public speech and monologue - public speech for S, and the three newspaper categories for W. The productivity is given as the number of hapaxes per million (part per million = ppm).

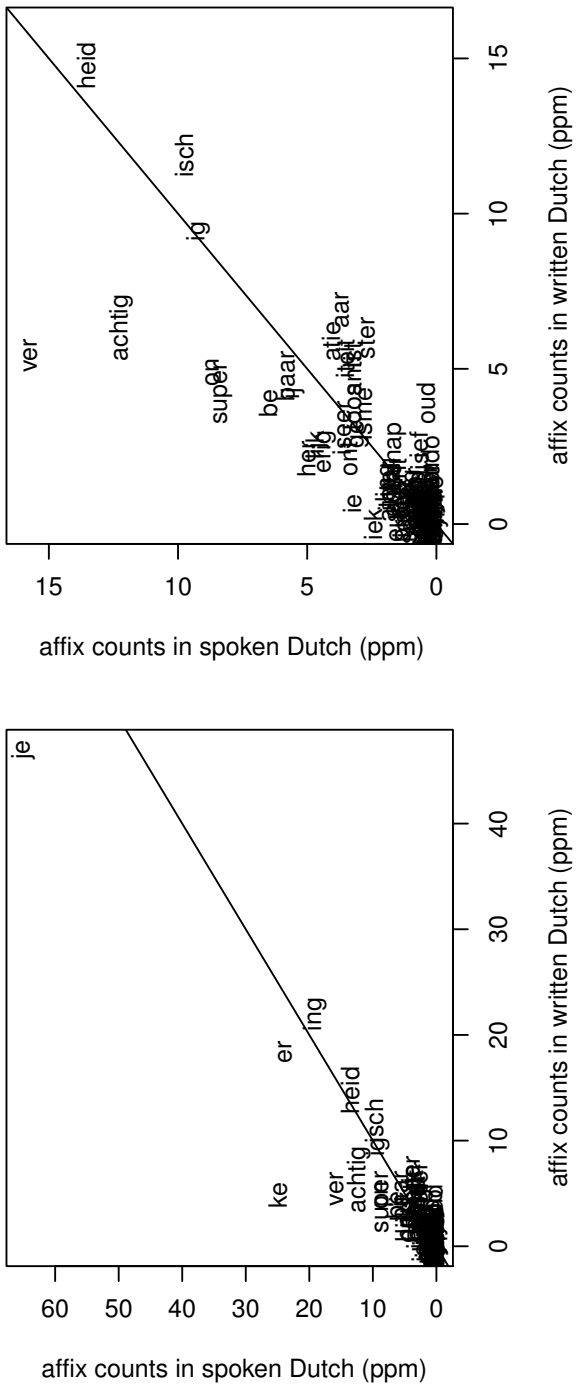


Figure 4.2: Affix productivity in written Dutch versus spoken Dutch for the separate affixes. The affix counts are presented as the number of hapaxes in million words (ppm). The right panel is a close-up of the lower left corner of the left panel.

Table 4.1: Analysis of deviance table for *derivational* productivity in the Newspaper corpus. Model: binomial, link: logit. Response: cbind(counts, 1500000 - counts). Terms added sequentially (first to last). Sign. codes: 0.001 '***', 0.01 '**'.

	df	deviance	resid. dev.	p	
NULL			3493.9		
country	1	0.7	3493.1	0.3870350	
register	2	94.3	3398.9	< 2.2e-16	***
affix	84	2809.1	589.7	< 2.2e-16	***
country: register	2	15.7	574.0	0.0003841	***
country: affix	84	120.7	453.3	0.0054282	**
register: affix	168	305.1	148.3	5.176e-10	***

4.3.1 Written Dutch

Derivational productivity

The next step was to analyze the distributional patterns of the hapax legomena in the written corpus. Their share in specific subcorpora was investigated by computing their logit values (the logarithm of their frequency divided by the number of the other tokens in the subcorpus). We decided to include the set of affixes in the statistical analysis as a fixed effect, for two reasons. First, our selection of affixes is not a random sample from the population of affixes, but the (almost) full selection of all affixes as documented in standard Dutch grammar. Second, even if one would argue that treatment of the affix factor as a random effect might be preferable, mixed-effects models assume random effects to be normally distributed. However, affix counts follow Zipfian distributions. As a consequence, mixed-effects models are not directly applicable (Keune et al., 2006).

We applied a generalized linear model with a binomial link. The results are given in Table 4.1. We started with the null model, containing no effects, and added new effects stepwise, starting with the three main effects. Significance was determined by the size of the deviance score (the difference between residual deviances of the preceding model and the model including the effect). Country was kept because of the strong two-way interactions in the final model. No further improvement was obtained by including a three-way interaction.

The final model selected contains all two-way interactions. The model selected revealed main effects for two out of the three predictors: register, and affix. The effect of register reflects the order visualized in Figure 4.1, the tabloids having the lowest derivational productivity and the quality newspapers having the highest one. Affix productivity varied from affix to affix: the three affixes with the highest overalll productivity were all suffixes, namely *-ing* ('tekening',

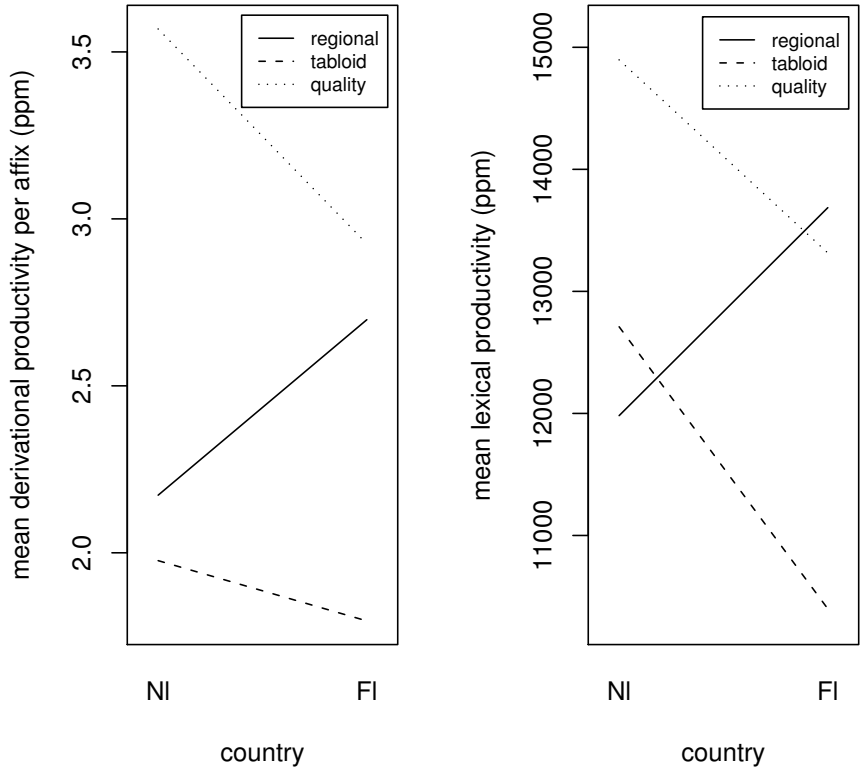
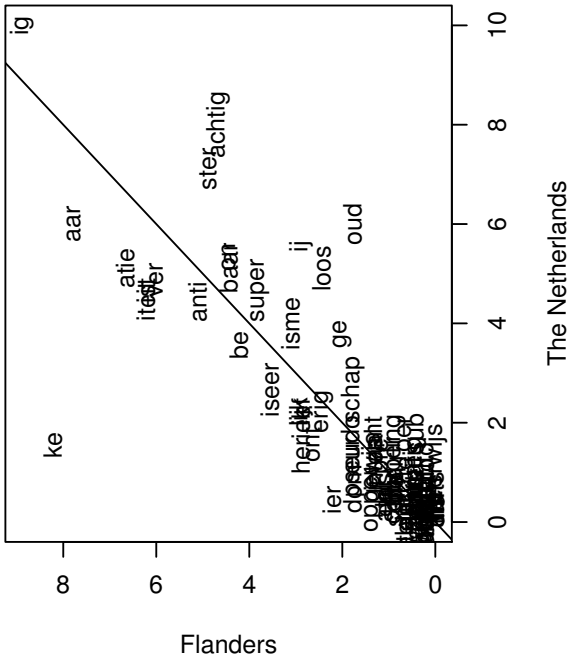


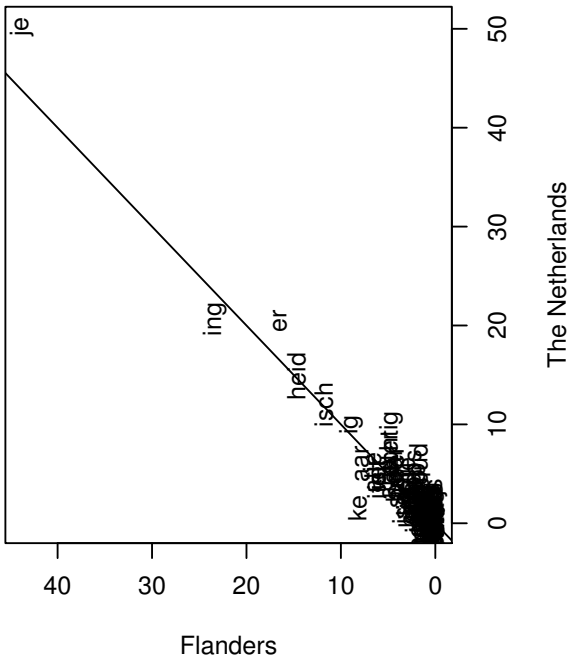
Figure 4.3: Interaction plots for derivational hapax legomena (left) and lexical hapax legomena (right) in the newspaper corpus.

'drawing'), *-er* ('zanger', 'singer'), and *-heid* ('slimheid', 'smartness'). These main effects were modulated by interaction effects.

The effect of register varies in relation to country. This interaction of country by register is visualized in the left panel of Figure 4.3 (the right panel on lexical productivity will be discussed later).

Figure 4.3 makes clear that the regional newspapers have a different position in the two countries. In the Netherlands affix productivity in the regional newspaper was more comparable to the tabloid, while it was more comparable to the quality newspaper in Flanders. This was confirmed by additional statistical tests. The Dutch quality newspaper had a higher productivity than the Flemish quality newspaper ($F(1, 168) = 8.1341, p < 0.0043$), while the regional newspaper had a higher productivity in Flanders ($F(1, 168) = 7.2428, p < 0.0071$). We come back to this outcome when discussing the results for lexical productivity.


ig
aar
ke
ste
achtig
atie
anti
super
be
isme
ge
loos
oud
ij
iseer
ier
Flanders
The Netherlands
0
2
4
6
8
10


je
ing
er
heid
isch
ig
ke
aar
Flanders
The Netherlands
0
10
20
30
40
50

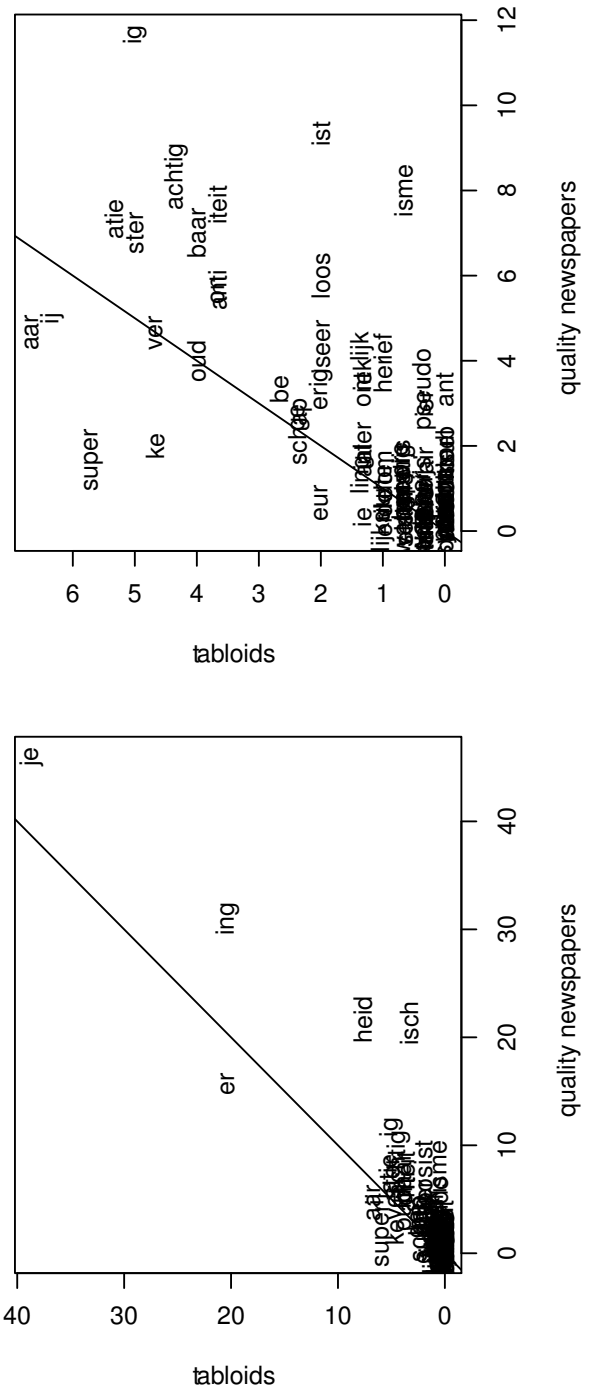


Figure 4.4: Affix productivity across register and country in parts per million (ppm). Right panels are close-ups of the lower left corners of the left panels.

Table 4.2: Analysis of deviance table for *lexical* productivity in the newspaper corpus. Model: binomial, link: logit. Response: cbind(counts, 1500000 - counts). Terms added sequentially (first to last). Sign. codes: 0.001 '***'.

	df	deviance	resid. dev.	p	
NULL			1430.21		
country	1	94.56	1335.66	< 2.2e-16	***
register	2	771.97	563.68	< 2.2e-16	***
country:register	2	563.68	1.452e-10	< 2.2e-16	***

The two other interactions relate to the different affixes. These interactions are best explored again by means of visualization. The upper panels of Figure 4.4 illustrate the interaction of affix by country. The frequencies of the derivational hapaxes are given in parts per million. The upper left panel shows that the diminutive suffix *-je*, which was excluded from the statistical analysis, enjoys the greatest degree of productivity in both the Netherlands and Flanders. Its location on the right of the line $Y = X$ (represented by the solid line) indicates that it is used more productively in the Netherlands than in Flanders. The next four most productive affixes, - *ing* ('handeling', 'action'), *-er* ('schrijver', 'writer'), *-heid* ('gekheid', 'craziness'), and *-isch* ('kritisch', 'critical'), exhibit a roughly equal productivity in the two countries. The upper right panel zooms in on the cluster of affixes in the lower left corner of the upper left panel. Affixes typically used in the Netherlands are *-ig* (' gelig', 'yellowish'), *-achtig* ('vergeetachtig', 'forgetful'), *-ster* ('schrijfster', 'female writer'), and *oud-* ('oud-directeur', 'former director'). The affix that is most typical for Flanders is the excluded Belgian Dutch diminutive suffix *-ke* ('danske', 'little dance').

The lower panels of Figure 4.4 illustrate the interaction of affix and register. We focused on the panels of national tabloids by quality newspapers as they produced the most outspoken results. The excluded diminutive suffix *-je* enjoys the greatest productivity, and occurs most frequently in the quality newspapers. The next three most productive affixes emerge as more productive in the quality newspapers: *-ing*, *-heid*, and *isch*. The differentiation for the less productive affixes is visible in the lower right panel. The vogue intensifier *super-*, ('supermooi', 'super beautiful'), occurs more frequently in the tabloids. The same applies to the excluded country specific diminutive variants *-ke* and *-ie*.

Lexical productivity

The next step was to analyse lexical productivity in the newspaper corpus. We ran a generalized linear model. Both the two main effects and their interaction were significant. Statistical details are given in Table 4.2.

The interaction of country by register is visualized in the right panel of Figure 4.3. As is evident from a comparison of the left and right panels of

Figure 4.3, the patterns for affix productivity and lexical productivity are remarkably similar. There is a cross-over pattern now for the regional newspapers. The Dutch regional newspaper has a lower score in the Netherlands than the tabloid. In Flanders the regional newspaper competes with the quality newspaper.

How similar are derivational and lexical productivity? We carried out a Pearson correlation test between the derivational and lexical productivity scores in the six newspapers involved. With a Pearson correlation of 0.91, this test confirmed the high similarity between the two types of productivity. The inclusion of the diminutive affixes slightly lowered the correlation to 0.90.

4.3.2 Spoken Dutch

Derivational productivity

The spoken data from the Spoken Dutch Corpus were analyzed in the same way as the written data from the CONDIV newspaper corpus. We had to exclude the public monologues for two reasons. First, they had a seperate status consiting of prepared speech. Second, the combination of the main effects gives twenty-four subcorpora. Twelve of the twenty-four subcorpora we obtained for monologues contained less that 5,000 words, which implies that not enough sociolinguistic information is available.

We applied a generalized linear model with a logit binomial link to the spoken data, including six predictors. Register covers private versus public dialogues. Affix was taken to be a fixed factor. As before, the analysis starts with the null model and is completed by adding effects, starting with the main effects and completed by adding all two-way interactions. No substantial three-way and higher interactions were found (explaining more the 5% of the sum of deviance scores of the main effects (7.3, when affix was not involved; 97.0 when affix was involved).

As shown in Table 4.3, a generalized linear model with all two-way interaction terms resulted in a model with main effects for five out of six predictors: register, gender, education level, age, and affix. As for the derivational productivity, no effect was found for country. Old Dutch male speakers with a high education level appeared to be the most productive age group. Middle aged speakers were more productive in the use of derivational affixes than young speakers. The remaining main effects will be discussed when interpreting the relevant interactions. We start with the two-way interaction where affix was not involved.

Five out of the ten interactions received significant p values. As for derivational productivity, we observe that the sum of the deviance scores of the five main effects involved (not affix) are larger. We calculated the sum of explained deviance for these main effects (147.6). We decided to interpret again only the interactions which explained more than 5% of the main effects involved in order to get the substantial effects (> 7.3), and which had a p value lower than 0.01

Table 4.3: Analysis of deviance table for *derivational* productivity in the Spoken Dutch Corpus. Model: binomial, link: logit. Response: cbind(counts, 1500000 - counts). Terms added sequentially (first to last). Sign. codes: 0.001 '***', 0.01 '**', 0.05 *'. edu = education level

	d	deviance	resid. dev.	p	
NULL			3721.7		
country	1	0.1	3721.6	0.717814	
gender	1	45.6	3676.0	1.440e-11	***
edu	1	29.9	3646.1	4.604e-08	***
age	2	29.9	3616.2	3.152e-07	***
register	1	42.1	3574.1	8.678e-11	***
affix	69	1792.8	1781.3	< 2.2e-16	***
country:gender	1	3.9	1777.3	0.046990	*
country:edu	1	1.7	1775.6	0.190466	
country:age	2	8.4	1767.2	0.014732	*
country:register	1	1.8	1765.4	0.177220	
gender:edu	1	0.1	1765.3	0.817760	
gender:age	2	1.3	1764.0	0.518808	
gender:register	1	6.1	1758.0	0.013854	*
edu:age	2	9.2	1748.7	0.009857	**
edu:register	1	2.6	1746.1	0.103978	
age:register	2	11.5	1734.5	0.003145	**
country:affix	69	130.9	1603.6	9.985e-06	***
gender:affix	69	101.5	1502.1	0.0066439	**
edu:affix	69	79.8	1422.3	0.1763139	
age:affix	138	210.5	1211.9	6.955e-05	***
register:affix	69	117.5	1094.3	0.0002445	***

in order to keep only the robust effects. Two interactions appeared to meet these two criteria: education by age and age by register. These interactions are visualized in the left panels of Figure 4.5. Old speakers with a high educational level appear to stand out as highly productive. Young speakers do not show a distinction between private and public speech, which may reflect an interesting age-bound effect, where young speakers are less productive as they are less experienced speakers or where young speakers produce more informal speech in the public domain.

The upper panels of Figure 4.6 visualize the interaction of affix by register. Specific affixes belong more to the private or more to public speech domain. While the diminutives, and the affixes *-achtig* and *-super* are typical for private dialogues, *-ing* and *-heid* are more typical for public dialogues.

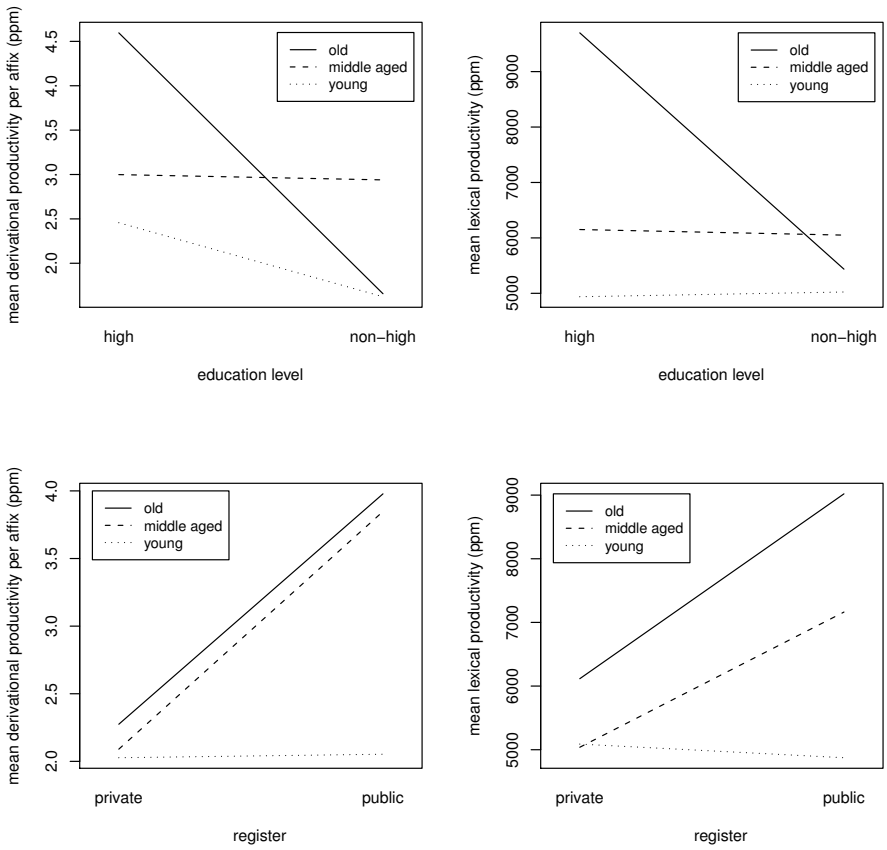
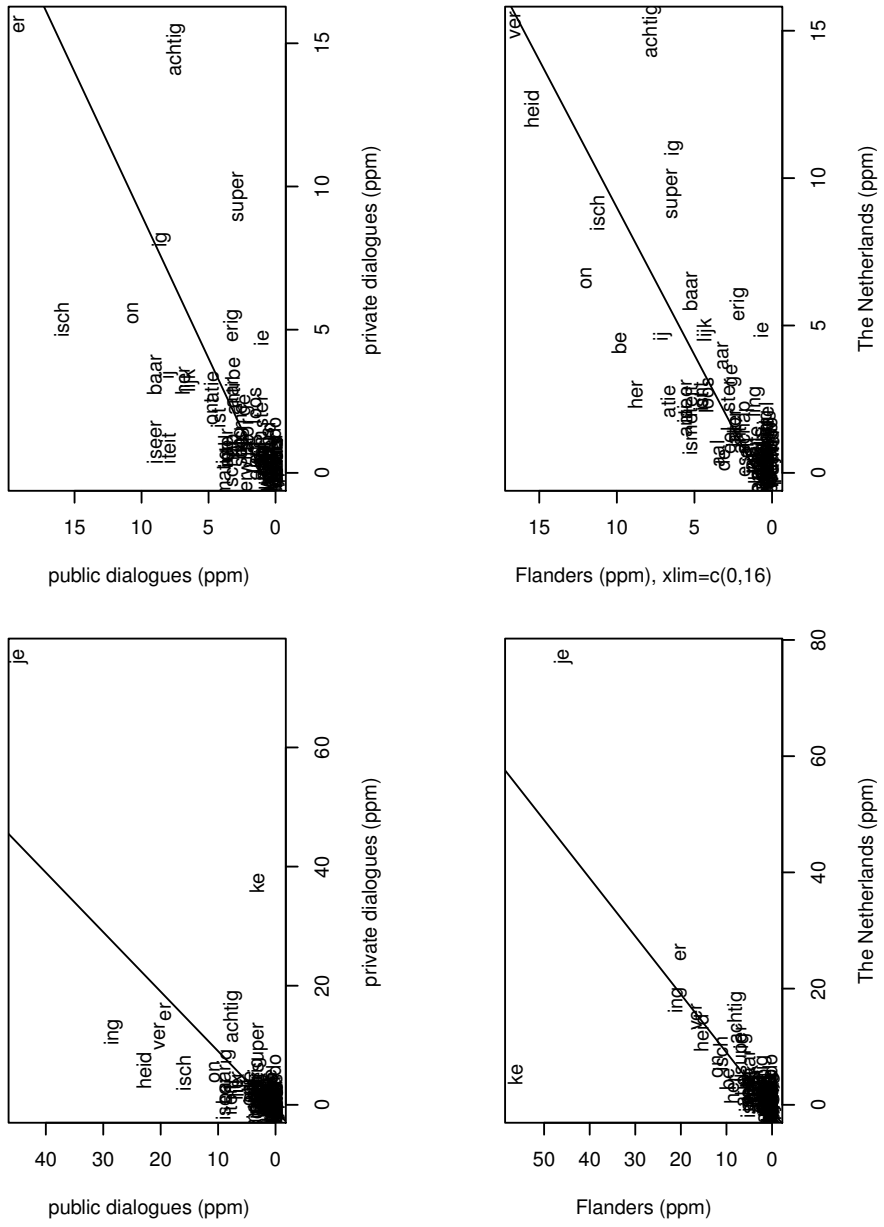


Figure 4.5: Interaction plots for derivational hapax legomena (left) and lexical hapax legomena (right) in the Spoken Dutch Corpus, for the interaction of education by age and the interaction of register by age


public dialogues (ppm)
private dialogues (ppm)
Flanders (ppm), xlim=c(0,16)
The Netherlands (ppm)
Flanders (ppm)
The Netherlands (ppm)

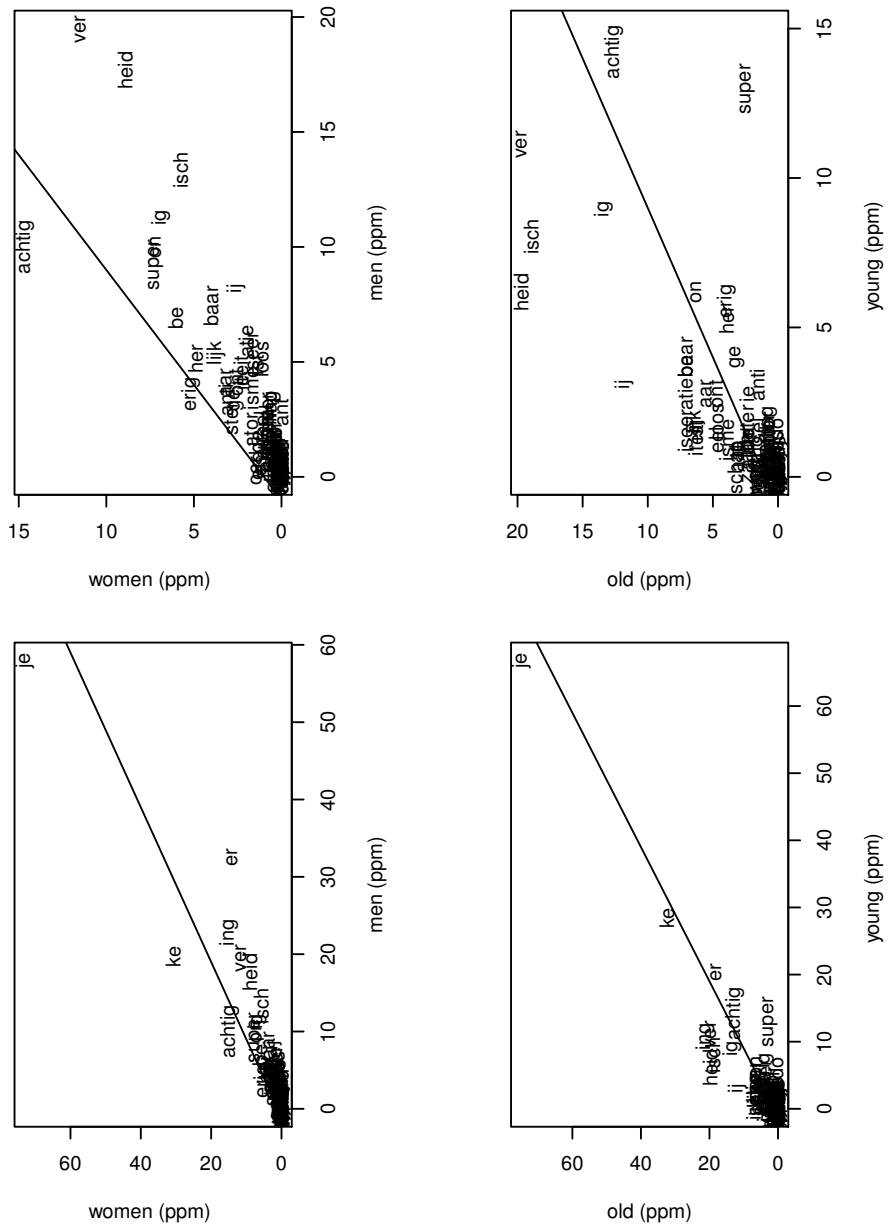


Figure 4.6: Affix productivity across register, country, gender, and age in parts per million (ppm). Right panels are close-ups of the lower left corners of the left panels.

The second row of Figure 4.6 visualizes the interaction of affix by country. The two most productive affixes, the diminutives *-je* and *-ke*, are as expected clearly differentiated in this plot: *-je* appears in the upper right corner, indicating that it is used more productively in the Netherlands, and *-ke* appears in the upper left corner, as it it is more productive in Flanders. The suffix *-er* is the most productive suffix typically for the Netherlands apart from the diminutives. We zoomed in on the cluster in the lower left corner, which resulted in the upper right panel. This panel shows us that the productivity of the suffixes *-ig* and *-achtig* and the prefix *super-* is higher in the Netherlands, while in Flanders the prefixes *her-, be-*, and *on-* are more productive. The suffix *-atie* was reported by Pauwels (1964) to be more productive in Flanders, and his conclusion is supported by our data: *-atie* is located to the left of the line $Y = X$, and *-ing* to its right.

The third row of Figure 4.6 visualizes the interaction of affix by gender. Men appear to use most affixes more productively then women. However, the diminutive affixes, are used more productively by women. The suffix *-achtig* ('vergeetachtig', 'forgetful'), is apart from the diminutives the most characteristic affix for women in the Corpus Spoken Dutch when we look at overpresentation.

Finally, the bottom panels of Figure 4.6 illustrate the interaction of affix by age. The right panel, which zooms in on the cluster in the lower left corner of the left panel, shows that older speakers use most affixes more productively than younger speakers: Most affixes are located to the left of the $Y = X$ line. The prefix *super-*, ('supermooi', 'super beautiful') is perhaps most characteristic for young speakers. This prefix orininates from the Latin word *super*, meaning for instance 'above', 'on top of' or 'beyond'. In the modern English language the prefix *super* obtained a highly general meaning, namely 'very' or 'extraordinary'. In the second half of the 20th century this way of using the prefix is taken over in the Dutch language. Productivity of the prefix has highly increased in the last decades, and explains the relatively high productivity among young speakers. They often use it where older speakers would make use of the standard intensifier (*heel*, 'very').

Older speakers prefer affixes such as *-ing* (*beslissing*, 'decision'), *-heid* (*mogelijkheid*, 'possibility'), *ver-* (*verbouwen*, 'renovate'), and *-isch* ('kritisch', 'critical'). Interestingly, the suffix *-lijk* (*gevaarlijk*, 'dangerous') also emerges as somewhat more productive for older speakers. Possibly, this difference reflects an ongoing language change. The suffix *-lijk* (old speakers: 6.3 ppm, young speakers: 1.9 ppm) is hardly productivity, as witnessed by the many high-frequency formations that regularly undergo substantial reduction (Keune et al., 2005). The present difference between age groups suggests that for younger speakers *-lijk* may even be less productive, which may be indicative of a diachronic process of affix attrition.

Table 4.4: Analysis of Deviance Table for *lexical* productivity in the Newspaper corpus. Model: binomial, link: logit. Response: cbind(counts, 1500000 - counts). Terms added sequentially (first to last). Sign. codes: 0.001 '***', 0.01 '**', 0.05 *'. edu = education Level

	df	deviance	resid. dev	p	
NULL			3234.9		
country	1	7.5	3227.3	0.00604	**
gender	1	735.4	2492.0	< 2.2e-16	***
edu	1	186.5	2305.5	< 2.2e-16	***
age	2	990.9	1314.6	<2.2e-16	***
register	1	400.0	914.6	< 2.2e-16	***
country:gender	1	37.6	877.0	8.607e-10	***
country:edu	1	33.2	843.8	8.329e-09	***
country:age	2	39.2	804.6	3.039e-09	***
country:register	1	26.5	778.0	2.582e-07	***
gender:edu	1	1.0	777.0	0.30700	
gender:age	2	7.6	769.4	0.02291	*
gender:register	1	3.0	766.4	0.08473	
edu:age	2	193.6	572.9	<2.2e-16	***
edu:register	1	44.6	528.2	2.374e-11	***
age:register	2	306.9	221.3	< 2.2e-16	***

Lexical productivity

The lexical productivity was analyzed in the same way as the derivational productivity. There were five predictors, as can be seen in Table 4.4. Again we expanded the model by incorporating all two-way interactions. Higher interactions did not pass the criteria of 5% explained variance (5% of 2320.3 = 161.7).

As in the analyses of written Dutch, we investigated overall lexical productivity in our subcorpora, in order to obtain a baseline against which morphological productivity can be evaluated. Again, we will only discuss those effects, explaining more that 5% of the total deviance in order to select only the robust effects. The main effects register, gender, education level, and age emerged as the main factors explaining an important part of the deviance. As with the derivational affixes, the highest lexical productivity was revealed by Old Dutch men with a high education level. Middle aged speakers showed more lexical productivity than Young speakers.

As with derivational productivity, the interactions of education level by age and age by register were the only interactions explaining more than 5% of the total deviance. These interaction are visualized in the lower panels of Figure 4.5. The main pattern of results for the lexical productivity is very similar to that

observed for derivational productivity. As with derivational productivity, old speakers show no more productivity as the other speakers, when they are not higly educated. The interaction of age by register only differs from derivational productivity in the higher productivity for old speakers compared to young and middle-aged speakers in private speech.

A Pearson correlation test confirmed the similarity between derivational and lexical productivity in spoken Dutch. We correlated the number of derivational and lexical hapaxes occurring in each of the 48 subcorpora we used in our analyses of spoken Dutch. The Pearson correlation was 0.84. After the removal of the four subcorpora in which no derivational hapax appeared, the correlation was even higher, namely 0.86. The inclusion of the diminutives to the derivational hapaxes, lowered the Pearson correlation to 0.78. The exclusion of subcorpora containing zero derivational hapaxes, resulted in a higher correlation of 0.85.

4.4 Conclusion and Discussion

In the present study we investigated variation patterns according the *use*, i.e. register, and *user* , i.e. the country factor and social factors, in derivational and lexical productivity in spoken and written Dutch. We studied whether derivational productivity goes hand in hand with lexical productivity or whether it is a separate form of productivity. We first analyzed derivational and lexical productivity in the main subcorpora of two large corpora, a written Dutch newspaper corpus (CONDIV and the Spoken Dutch Corpus (CGN). We looked for affix-specific effects in derivational productivity to establish whether the derivational productivity patterns we found are global overall effects pertaining to all affixes or whether there are substantial affix-specific differences, implying that affixes are primarily marked by their own productivity.

The CONDIV newspaper corpus contains newspapers from the Netherlands and Flanders (country). Within these countries we compared three newspaper registers, namely quality newspapers, tabloids, and regional newspapers. In the Spoken Dutch Corpus we made a distinction between the Netherlands and Flanders (country), and contrasted the following speech registers: formal monologues, formal dialogues, and informal dialogues (spontaneous speech). In the Spoken Dutch Corpus social meta-data about the speakers was available. We therefore also studied the effects of gender, education level, and age. Productivity was established on the presence of hapax legomena, words or derivations occurring only once in a corpus. We applied generalized linear models with a binomial link function (GLM) to analyze the distribution patterns of the hapax legomena. Previous research Keune et al. (2006), testified that these models produce robust results, even when the probability of a success (a specific word is a derivational hapax legomenon) in the data is extremely small compared to the probability of a failure (a specific word is not a derivational hapax legomenon). Furthermore, generalized linear models proved to be highly

suitable for handling data with highly varying cell sizes.

Our expectation that derivational productivity mirrors general lexical productivity turned out to be correct. The overall analyses returned a strong correlation between derivational and lexical productivity in different newspaper and speech registers. Also the outcomes of the GLM analyses in which we subdivided the data according to newspaper and speech register, country and the three social factors (age, gender, education) in spoken Dutch registers, demonstrated a clear and robust correspondence between derivational and lexical productivity. All main effects of the two productivity measures pointed in the same direction, and even the interactions patterns were similar in most respects. Only the probability values of the effects for lexical productivity were often lower than those for derivational productivity, but this is a frequency effect. The number of derivational hapaxes in the data was much lower than the number of lexical hapaxes, which causes data sparseness and therefore gives more uncertainty and less power in the analyses. Our research made clear that the lexical productivity of a (sub)corpus is for only a very small part the outcome of the presence of derivationally formed neologisms. There are many other processes involved in the creation of neologisms. More substantial word producing processes that allow new words to enter the lexicon are for instance word compounding, word composition and word borrowing.

In line with previous research from Biber and Conrad (2009) and Plag et al. (1999), we found a higher lexical productivity for written than for spoken language, and a higher productivity in more formal registers. There was one remarkable result: the high hapax ratio for public monologues. As we already noted in the result section, this high ratio may be partly due to the relatively low number of monologue data available and to a higher resemblance between spoken monologue data (more prepared, sometimes on paper) and written registers. Furthermore, the available monologue data have a social bias, as by far the most monologue data comes from higher educated speakers.

As for derivational productivity, however, spoken Dutch was not less productive than written Dutch. This high degree of productivity is partly due to the diminutive suffixes. Removing them, however, did not result into a much higher productivity figure for written Dutch. In most analyses the diminutives came out as a deviant member of the set of affixes. Because of their frequent occurrence they highly influenced the results. In the result section we already mentioned the special status of the diminutives because of their vast domain of potential applications (even more than all nouns), and their high degree of productivity and their variability in semantics (not only indicating smallness, but all kinds of meanings related to smallness, like affectionate meanings). We excluded the diminutives from our analyses, but kept them in visualizing the properties of the different affixes, to provide information on the distributional properties of the diminutive suffixes.

The outcome that part of the affixes were more productive in spoken Dutch while other affixes were more productive in written Dutch, seems not to be

adequate enough to explain fully that affix productivity in spoken Dutch was not much lower than in written Dutch. An interesting explanation is that in spontaneous speech, speakers make active use of the productive properties of affixes to coin new words. While the frequency of complex words is higher in written than in spoken language, complex words used in the written registers may more often be already well-established words. In spontaneous speech there is not much time to reflect on word choice and to reconsider what words can be produced best. Time pushes and productive affixes can be a help in making lexical decisions. Since spontaneous speech commonly is a more informal register or style, it is to be expected that the affixes that are most productive in spontaneous speech are the more informally used affixes. It is therefore not a surprise that the diminutive suffixes, which fit better in the involved language registers, occur highly frequent in spontaneous speech. This interpretation is strengthened by the observation that groups of speakers for which we found a relatively low level of derivational productivity for most affixes, like women's speech and informal speech, showed an increased productivity for the diminutives.

The high impact of the diminutive is also shown by the diminutive variant *-ke*, which is typically used in Flanders. Because of its country specific character and its frequent occurrence, this variant caused an effect for country all by itself. The conclusion is that investigating derivational productivity always requires scrutinizing the role of all individual affixes before any general conclusions can be drawn.

Not only the diminutives showed variation in their frequency of use among the different registers. The registers with the highest overall affix productivity showed the highest productivity figures for most affixes. However, some affixes were more typical of a specific register or related to a specific social variable. The suffix *-achtig* (*vergeetachtig*, 'forgetful'), for instance, is an affix characteristics of women, and the prefix *super-*, (*supermooi*, 'super beautiful') is characteristic of younger speakers.

In both written Dutch and spoken Dutch, register emerged as an important predictor of the degree of productivity. The more formal a register was, the higher its derivational and lexical productivity. This finding is consistent with the informational versus involved' dimension that Biber identified to contrast written and spoken language, and also to characterize different registers within written and spoken language (Biber, 1988, 1995; Biber et al., 1998; Biber and Conrad, 2009).

Country only emerged as an important factor in lexical productivity in written language. Even in this situation it proved to be register specific: the quality newspaper and the tabloid have higher lexical productivity figures in the Netherlands than in Flanders, while the regional newspaper is more productive in Flanders than in the Netherlands. This can perhaps be explained by the aspirations of the Flemish regional newspaper the to be a quality newspaper.

The absence of a global productivity difference between the Netherlands

and Flanders, corresponds to earlier findings in Keune et al. (submitted). They did not find a productivity effect for country in spontaneous speech either, but they found lexical differences between the two countries in the occurrences of adjectives, verbs and most common words. However, non of these effects were typical of the word class category as a whole. All effects were word-specific. Closer inspection of the affix-specific variation in productivity in the present paper confirmed that also on the derivational level, the differences between countries were affix-specific.

Each of the social factors included in the analyses of spoken Dutch turned out to be influential. We found that old men with a high education level used both derivational and lexical items most productively. The higher productivity for men is in agreement with the higher male Type-Token Ratio revealed by Härnqvist et al. (2003), Van Gijsel et al. (2006) and Van Gijsel (2007), and the higher male lexical productivity found by Keune et al. (submitted). From the last study it appeared that lexical effects for gender are global. This is in line with our results. The removal of a number of gender specific affixes did not influence the outcomes of our analyses. Only three affixes had a higher productivity for women, namely the diminutives *-je* and *-ke*, and the suffix *-achtig*). This raises the question if and how this global gender productivity effect can be related to the systematic gender effect found in sociolinguistic studies on language variation and change (Coates, 1998; Newman et al., 2008; Härnqvist et al., 2003).

Our finding that highly educated speakers reveal a higher derivational and lexical productivity, is in line with previous research on effects of education level. Härnqvist et al. (2003) found a higher Type-Token Ratio for highly educated speakers. Keune et al. (2005) showed that highly educated speakers use more words ending in the suffix *-lijk* than non highly educated speakers. This outcome also matches the high productivity in quality newspapers, since they aim at a higher educated readership.

The degree of productivity turns out to increase with the speaker's age. It seems to show that skills necessary to create new words and word forms, are developed over a longer time span. However, the age effect is not present among non-highly educated speakers. It means that a higher productivity does not develop automatically, but needs to be triggered. The absence of an effect of age in private dialogues is not surprising. We already mentioned that private, non-professional spontaneous speech is expected to be a more involved style. It is therefore likely that highly educated speakers do not fully exploit all their derivational and lexical resources in more involved speech. This result explains why Keune et al. (submitted) did not trace any educational effects in their corpus of spontaneous speech, while Härnqvist et al. (2003) and Keune et al. (2006) both observed a higher lexical richness for highly educated speakers, as their data contained speech styles other than private, spontaneous speech.

There turn out to be factors that have an impact on the productivity found in a text, including register differences and sociolinguistic factors (country,

social variables). Our study on affix-specific variation also disclosed that there is not something like an ideal situation which evokes a maximal productivity of all individual affixes at the same time. We expect that this affix and word-specific finding generalizes to other word producing processes. Given the resemblance of overall lexical and derivational productivity, it seems evident to suppose that other word producing processes are submitted to the same set of sociolinguistic constraints and register bound limitations.

References

Anshen, F. and M. Aronoff, 1997. Morphology in real time. In G. E. Booij and J. van Marle, eds., *Yearbook of Morphology*. Kluwer Academic Publishers, Dordrecht, 9–12

Baayen, R. H., 1994. Derivational productivity and text typology. *Journal of Quantitative Linguistics*, 1: 16–34

Baayen, R. H., 2009. Corpus linguistics in morphology: morphological productivity. In A. Luedeling and M. Kyto, eds., *Corpus Linguistics. An international handbook*. Mouton De Gruyter, Berlin, 900–919

Baayen, R. H. and A. Neijt, 1997. Productivity in context: a case study of a Dutch suffix. *Linguistics*, 35: 565–587

Baayen, R. H., R. Piepenbrock and L. Gulikers, 1995. *The CELEX lexical database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA

Baayen, R. H. and A. Renouf, 1996. Chronicling The Times: Productive Lexical Innovations in an English Newspaper. *Language*, 72: 69–96

Biber, D., 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge

Biber, D., 1995. *Dimensions of Register Variation*. Cambridge University Press, Cambridge

Biber, D. and S. Conrad, 2009. *Register, Genre, and Style*. Cambridge University Press, Cambridge

Biber, D., S. Conrad and R. Reppen, 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, Cambridge

Chesley, P. and R. H. Baayen, 2011. Predicting new words from newer words: Lexical borrowings in French. *Linguistics*, 48: 1343–1374

Coates, J., ed., 1998. *Language and gender: A Reader*. Blackwell, Oxford

Geerts, G., W. Haeseryn, J. de Rooij and M. C. van den Toorn, 1984. *Algemene Nederlandse Spraakkunst.* Wolters-Noordhoff, Groningen

Gries, S. T., 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement.* Continuum International Publishing Group Ltd., New York

Grondelaers, S., K. Deygers, H. van Aken, V. van den Heede and D. Speelman, 2000. Het CONDIV-corpus geschreven Nederlands. *Nederlandse Taalkunde*, 5: 356–363

Härnqvist, K., U. Christianson, D. Ridings and J.-G. Tingsell, 2003. Vocabulary in interviews as related to respondent characteristics. *Computers and the Humanities*, 37: 179–204

Keune, K., M. Ernestus, R. van Hout and R. H. Baayen, 2005. Social, geographical, and register variation in Dutch: From written 'mogelijk' to spoken 'mok'. *Corpus Linguistics and Linguistic Theory*, 1: 183–223

Keune, K., S. van Gijsel, R. van Hout and R. H. Baayen, submitted. Sociolinguistic patterns in dutch: Measuring lexical characteristics of spontaneous speech. Speech communication

Keune, K., R. van Hout and R. H. Baayen, 2006. Socio-geographic variation in morphological productivity in spoken dutch: A comparison of statistical techniques. In J.-M. Viprey, ed., *Actes des 8es journées internationales d'analyse statistique des données textuelles*, volume 2. Presses Universitaires de Franche-Comté, 571–580

Newman, M. L., C. J. Groom, L. D. Handelman and J. W. Pennebaker, 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45: 211–236

Oostdijk, N. H. J., 2002. The Design of the Spoken Dutch Corpus. In P. Peters, P. Collins and A. Smith, eds., *New Frontiers of Corpus Research.* Rodopi, Amsterdam, 105–112

Pauwels, J. L., 1964. Woorden op -atie en -ering in het nederlands. *Verslagen en Mededelingen van de Koninklijke Vlaamse Academie voor Taal- en Letterkunde*: 205–210

Plag, I., C. Dalton-Puffer and R. H. Baayen, 1999a. Morphological productivity across speech and writing. *English Language and Linguistics*, 3 (2): 209–228

Plag, I., C. Dalton-Puffer and R. H. Baayen, 1999b. Productivity and register. *Journal of English Language and Linguistics*, 3: 209–228

Savieckienė, I. and W. U. Dressler, eds., 2007. *The Acquisition of diminutives. A cross-linguistic perspective.* John Benjamins, Amsterdam, New York

Smith, J. A. and C. Kelly, 2002. Stylistic constancy and change across literary corpora: Using measures of lexical richness to dat works. *Computers and the Humanities*, 364: 411–430

Tagliamonte, S., 2008. So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics*, 12: 361–394

Van den Bosch, A. and W. Daelemans, 1999. Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99*. University of Maryland, USA, 285–292

Van Gijsel, S., 2007. Sociovariation in Lexical Richness. A Quantitative Corpus Linguistic Analysis. Ph.D. thesis, Katholieke Universiteit Leuven

Van Gijsel, S., D. Speelman and D. Geeraerts, 2005. A variationist, corpus linguistic analysis of lexical richness. In *Corpus Linguistics 2005*. Birmingham

Van Gijsel, S., D. Speelman and D. Geeraerts, 2006. Locating lexical richness: A corpus linguistic, sociovariational analysis. In *Proceedings of JADT 2006*. Besancon: Université de France-Comte, 961–971

CHAPTER 5

Sociolinguistic patterns in Dutch: Measuring lexical characteristics of spontaneous speech[1]

Abstract

This study addresses the existence and origins of sociolinguistic variation in the lexical spectrum of spontaneous speech in the Spoken Dutch Corpus. Three types of lexical measures were applied, lexical diversity (types, hapaxes), lexical density (nouns, adjectives, verbs) and lexical communality (most common words). A comparison of random and non-random text samples showed that random samples outperformed the non-random samples in stability and power. Principal Components Analysis was used to obtain an overview of the global variation patterns in the multivariate distribution of the six lexical measures and four social variables (country, gender, age, and education level). Linear models were applied to unravel the impact of the social variables in more detail. Significant lexical variation emerged for all four social variables. Most effects appeared to be word-specific, only the effects of gender were global and systematic, suggesting a pervasive style difference between women's and men's speech.

Keywords: lexical variation, corpus analysis, sampling methods, spontaneous speech analysis, speaker characteristics, lexical richness.

[1] This study, co-authored by Sofie van Gijsel, Roeland van Hout and Harald Baayen, has been submitted for publication.

5.1 Introduction

Spontaneous speech fragments show a broad and varying spectrum of lexical elements. Some fragments contain many different words, whereas others are marked by the occurrence of highly frequent function words or by a high number of verbs, nouns or adjectives. This invites the question: is this lexical spectrum somehow systematically related to external, sociolinguistic variables? In the literature, much research has been carried out in sociolinguistics on phonological, morphological and syntactic patterns of variation, but broader patterns of lexical variation have received little attention.

Broader lexical measures have been successfully used to explain variation in stylometric research, register variation studies, authorship attribution studies, and acquisition research. In stylistic research on language corpora, Biber (1988, 1995), and Biber and Conrad (2009) identified seven dimensions of language variation between registers of which the first dimension, most suitable to contrast written and spoken language, was 'informational' versus 'involved'. The use of nouns, long words, prepositions, and a high Type-Token Ratio (TTR) appeared as most typical for informational production, whereas private verbs[2], that-deletions, present tense verbs, and second person pronouns were most typical for involved production. Baayen (1994) and Plag, Dalton-Puffer and Baayen (1999) demonstrated that derivational affixes in general were used to a different extent in spoken and written language, with an obviously higher frequency in written language. Van Gijsel, Speelman and Geeraerts (2005, 2006) showed differences in lexical richness of the Dutch language between various registers of spoken language, the more formal registers having a higher richness. Burrows (1992a,b, 1993a,b) revealed regional variation, diachronic change and gender-specific differences in literary studies on the basis of the most common words (the highest frequency words). In the field of authorship attribution, Holmes (1994); Baayen, Van Halteren and Tweedie (1996), and Baayen, Van Halteren, Neijt and Tweedie (2002) showed that individual language users can often be distinguished by lexical and syntactic characteristics such as the most frequent words, the number of hapax legomena (words occurring only once) in a text, and Part Of Speech (POS) n-grams.

In acquisition research, global measures of lexical richness were applied to quantify size and diversity of a vocabulary and to index lexical development on the basis of spontaneous speech (Laufer and Nation, 1995). Vermeer (2000) examined the value of a series of lexical richness measures on Dutch L1 and L2 children, and concludes that the Guiraud measure gives the best results. Malvern and Richards (2002) propose an alternative to the standard lexical measures, VOCD (see CHILDES, the CLAN programs), a measure that does not work in comparing the vocabularies of Dutch L1 and L2 school children (Van Hout and Vermeer, 2007).

[2]Biber's definition of private verbs: *'Private verbs express intellectual states (e.g.,'believe') or non observable intellectual acts (e.g. 'discover')'* (Biber, 1988: 242).

Studies on variation in the domain of more general lexical characteristics mostly cover written data and corpora. Newman, Groom, Handelman and Pennebaker (2008) for instance, analyzed 14.000 written text samples with a series of lexical properties to determine if gender had an impact on text differences, and found many significant differences between men and women. Argamon, Koppel, Fine and Shimony (2003) explored lexical syntactic differences between male and female writing in a range of formal genres from the British National Corpus. They found gender differences in the use of pronouns in particular, and in certain types of noun modifiers. Research addressing sociolinguistic effects in lexical variation in spontaneous speech is quite limited. Rayson, Leech and Hodges (1997) investigated frequency differences for a large number of keywords in conversations from the British National Corpus. They defined speaker groups by gender, age and social class, and revealed keyword differences for each of these speaker groups. Furthermore, they compared written language with speech and found, in agreement with the results of Biber (1988, 1995) and Biber and Conrad (2009), that written language was more informational and speech more involved.

Härnqvist, Christianson, Ridings and Tingsell (2003) explored variation in vocabulary variables such as the Type-Token Ratio, the number of hapax legomena, word length, and POS across gender, socio-economic group, and education level, on the basis of 415 interviews with Swedish men and women. They found differences in vocabulary richness and POS for gender and education level. The number of hapax legomena appeared to be the best measure for richness. Keune, Ernestus, Van Hout and Baayen (2005), and Keune, Van Hout and Baayen (2006) showed sociolinguistic variation across speech and writing in both morphological and lexical productivity, and in the use of the most common words. Van Gijsel et al. (2005, 2006) included sociolinguistic variables in their research on lexical richness in spontaneous speech, dialogues and monologues. Register differences emerged as the factor explaining most variation. However, the speaker's country, gender, and age emerged as influential factors of richness too.

In this study we want to investigate whether general lexical characteristics show sociolinguistic variation in spontaneous speech, by systematically including external, social variables. If we can prove the relevance of external variables, the domain of sociolinguistic variation studies should be extended to include general lexical measures. The Spoken Dutch Corpus (CGN) (Oostdijk, 2002) is a spontaneous speech corpus that has a rich social stratification. It contains speaker information on country (the Netherlands vs. Flanders), age, education level and gender. We selected this corpus to investigate the social stratification a set of global lexical characteristics and consider what global lexical measures to include in this investigation.

A self-evident lexical measure is lexical diversity. An extensively studied and frequently applied measure for lexical richness or diversity (or vocabulary richness) is the Type-Token Ratio (TTR; Tweedie and Baayen, 1998; Arnaud,

1984; Richards, 1987). This measure counts the number of different word forms and word types in a text, and divides the number of word types by the number of word forms. Another measure used to determine the richness or creativity of a text is the number of hapax legomena (Baayen, 2001), i.e. the number of words occurring only once in a text. Smith and Kelly (2002) used this measure in a stylometric study as a distinguishing characteristic of an author's style. In morphological research, the rate of hapax legomena gives insight in the morphological productivity of a text (Keune et al., 2006).

The lexical density of a text, measured by the number of content words occurring in that text, is a measure frequently used in applied linguistics in order to determine a speaker's fluency in a second language. Johansson (2008) compares diversity and density in Swedish speech and writing in a developmental perspective. Heylighen and Dewaele (2002), however, showed that content words like verbs, nouns and adjectives, do not behave similarly. There is a difference between Parts Of Speech (POSs) that tend to increase with the formality of a text and POSs that tend to increase with the contextuality of a discourse. The POSs increasing with the formality of a text include nouns, adjectives, prepositions and articles, and the POS likely to increase with the contextuality of discourse consists of pronouns, verbs, adverbs, and interjections.

The number of most common (highest frequency) words (MCWs) in a text measures the lexical communality of a text. This measure appeared to be a good discriminator for authorship attribution (Burrows, 1992a, 1993a) as well as to discriminate sociolinguistic differences (Keune et al., 2005). The 30 or 50, or for instance, 80 most frequent words of a corpus are used as the most common words. The MCWs mainly comprise closed-class function words such as conjunctions, pronouns, prepositions, and determiners. Variation in the use of MCWs, tends to represent variation in syntactic habits (Baayen et al., 1996).

Given the earlier outcomes on lexical variation, we decided to include three types of lexical measures:

I. Lexical diversity: types (= TTR, given the fixed sample size), hapax legomena

II. Lexical density: nouns, adjectives, verbs (tokens)

III. Lexical communality: most common words (MCWs) (tokens)

What external effects can we expect from previous research? Gender differences in language use have been studied extensively (Coates, 1998). Most findings are consistent with the 'informational versus involved' dimension that Biber (1988, 1995) identified. Härnqvist et al. (2003) revealed a higher Type-Token Ratio for men. Rayson et al. (1997) found that male speakers favor common nouns, whereas female speakers prefer proper nouns, personal pronouns, and verbs. Newman et al. (2008) found that female language included more pronouns and social words, a wide variety of other psychological references, and more verbs. Male speech, on the other hand, was more likely to

include long words, numbers, articles, and prepositions. On the basis of these results we expect to find a higher lexical richness for men, and a higher lexical communality for women. Furthermore, we expect differences in the behavior of the POSs measuring the lexical density. In line with previous research, we expect nouns to be typical for male speech, and verbs for female speech.

A second external factor studied is the difference between Dutch as spoken in Flanders versus the Netherlands. While in the Netherlands there was a normal, continuous standardization process of the Dutch language, the standardization of the Dutch language in Flanders was politically hindered from the 16th to the 19th century. In the 19th century Belgian language planners decided to adopt the more prestigious Netherlandic Dutch as the standard language. It was recognized as an official language, alongside French in 1898. Only after 1930 Dutch did become the single official language in Flanders. Written Belgian Dutch has converged with written Netherlandic Dutch. However, this is not the case to the same extent for spoken Belgian Dutch (for a detailed overview of the situation of the standard Dutch in the two countries, see Grondelaers and Van Hout, 2011). Over the years, a Flemish version of standard Dutch emerged. This Flemish variety of standard Dutch is reserved for more formal situations (Jaspaert, 1986; Van den Toorn, Pijnenburg, Van Leuvensteijn and Vanˆder Horst, 1997; Geeraerts, 2001; Geeraerts, Grondelaers and Speelman, 1999). In less formal situations 'Tussentaal' ('in-between language', between standard and dialect) is used. This is a supra-regional variety of Belgian Dutch that retains characteristics of Flemish dialects and the standard language. There are clear differences between the spoken variety of standard Dutch, often referred to as VRT-Dutch[3], and Tussentaal. However, readjusting a few characteristics of this Tussentaal as spoken in for instance animated cartoons, would result in a variety of Dutch strongly overlapping with VRT-Dutch (Grondelaers and Van Hout, 2011). Except for a number of specific and highly frequent characteristics of Tussentaal, Dutch as spoken in Flanders is not at all that different from Dutch as spoken in the Netherlands. We therefore do not expect to find global lexical differences between the two speech communities. However, previous research comparing Dutch from the Netherlands to Dutch from Flanders has shown that there are non-global, word-specific differences in the lexicon (Geeraerts et al., 1999), but at the same time there is lexical convergence, even at the word-specific level (Geeraerts et al., 1999; Grondelaers and Van Hout, 2011).

With respect to age, we expect a higher lexical richness with older speakers. This is in agreement with the higher lexical diversity for adults in Johansson (2008), and the finding of Keune et al. (2006) that young Dutch female speakers with a non-high education level revealed the lowest degree of lexical creativity in the Spoken Dutch Corpus. With respect to lexical density, we expect to find

[3]Vlaamse radio en Televisie (Flemish Radio and Television) The term 'VRT-Dutch' refers to the important role of the VRT in the propagation and diffusion of this spoken variety of standard Dutch (see Vandenbussche 2010)

that old speakers use more nouns. Older speakers seem to have the advantage of having expanded their lexical knowledge over a longer time span, including all lexical innovations and names for new products and tools. Furthermore, Rayson et al. (1997), find that young speakers make intensive use of some specific adjectives. This finding might also be reflected in the present data.

Finally, given the results of Härnqvist et al. (2003), and Keune et al. (2006) who both find a higher lexical richness for highly educated speakers, we expect to find the same results in the present research, although there is a good reason to assume that the differences will be modest. The data we will analyze come from private spontaneous speech, as we wanted to investigate informal speech, the register usually investigated in sociolinguistics. We expect that highly educated speakers do not fully exploit all their lexical resources in private, spontaneous speech.

How should we proceed to document potential lexical social stratification in the CGN? The most obvious choice seems to be to use all the data available. Measures of lexical richness, however, have two major drawbacks. First, they are highly text-length dependent: the longer a text the lower the TTR (Tweedie and Baayen, 1998; Vermeer, 2000), and the lower the relative number of hapaxes legomena (Baayen, 1996). The other drawback is the impact of topic dependency. Increasing the number of topics has the immediate effect of higher outcomes of the two measures. There is a strong awareness of this problem in the field of word frequency distribution modeling (Baayen, 1996, 2001). Evert and Baroni (2005) demonstrate thematic bias reduction, or even removal, by randomizing tokens in corpora. They automatically predict growth curves for the number of hapax legomena occurring in a corpus with the help of Large Number of Rare Events models, and show more precise predictions when the tokens are randomized.

To cope with this problem we will use non-random and random samples in our research on the CGN, as explained in the method section. In addition, we will work with samples from the CGN corpus, to include the degree of variation in the subcorpora, as explained again in the method section. We want to demonstrate that a random sampling approach is the best way to deal with larger speech corpora that can be stratified in subcorpora, on the basis of external, social factors.

We start the analysis of the data samples with a factor analysis to investigate how different our six lexical measures are related to each other and to the sociolinguistic factors we investigate: country, gender, education level, and age. In theory, there may be a great overlap between the measures, as can be expected for the number of types and the number of hapax legomena, both being measures of lexical richness. Biber (1988, 1995) has successfully applied factor analysis regularly to explore the number of different dimensions in the many text measures he explored. Our next step in the analysis is to apply linear model analysis including the four sociolinguistic factors as the independent variables and the six lexical measures as the dependent variables, to investigate

the strength of these factors as well a the strength of their interactions. Different measures turn out to reveal completely different sociolinguistic patterns. Finally, we will take a closer look at our data to determine whether the variation patterns observed are global or word-bound. We will discuss our approach and the results in the final section and conclude with some remarks for further research.

Method

In order to explore the sociolinguistic variation patterns in spoken Dutch, we selected the components from the Spoken Dutch Corpus (CGN) containing private spontaneous speech. The total corpus contains approximately 9 million words and comprises a large number of samples of (recorded) text, amounting to about 800 hours of speech. Two thirds of the corpus were collected in the Netherlands and one third in Flanders, the Dutch-speaking, northern part of Belgium. The entire corpus is orthographically transcribed, lemmatized and tagged for POS. The corpus is structured along 15 registers or components, ranging from very informal face-to-face conversations to more formal components, such as lectures and seminars and even read-aloud speech. We selected the spontaneous speech data obtained in private, non-professional telephone and face-to-face conversations (the corpus components A, C, and D). This part of the CGN corpus comprises approximately 4.7 million words (cf. Oostdijk, 2002).

The CGN corpus contains detailed speaker information. We stratified our spontaneous speech corpus by applying four speaker characteristics or variables: country (the Netherlands versus Flanders), education level (high versus non-high)[4] gender, and age (young: < 40; middle aged: 41–60; old: > 60). The result is $2 \times 2 \times 3 = 24$ strata or subcorpora. These subcorpora varied substantially in size: from 10,966 words of old Flemish non-highly educated men, to 724,811 words of young Dutch highly educated women.

The standard procedure in corpus linguistics is to treat complete subcorpora as the units of analysis, an approach that is appropriate in many cases, especially when specific linguistic constructions or variables are being investigated. However, we want to apply global or overall measures. Such measures can be calculated for subparts of the available subcorpora as well (subsampling the sample), having the advantage that information can be obtained about the variability (perhaps it is better to say the stability) of the characteristic investigated. Another consideration in favor of choosing an alternative sample design is that some of the measures we apply are text length dependent and that applies particularly to lexical richness measures like the TTR and the number of hapax legomena. A straightforward solution is taking the size of the smallest subcorpus as the yardstick, reducing the larger subcorpora accordingly. But

[4]High: attended bachelor or master education

there is an obvious disadvantage: large parts of the larger subcorpora are in fact discarded and do not play any role any longer in determining the outcomes of the measures. Sampling the subcorpora repeatedly is a better solution (Baayen, 2001), since it gives additional information on the stability of the measures in the subcorpora, by computing the variance of the outcomes. Sampling from a sample is a well-known procedure in statistics, the most famous examples being bootstrapping and randomization tests (e.g., Wilcox 2010). It is not obvious however, how large the subsamples should be to obtain an optimal mix between frequency of resampling and sample size. For the present study, the subsamples should be large enough to give information on the impact of the speaker variables we want to investigate. We decided to estimate the optimal sampling size (the number of word tokens) by computing the number of types for different token sizes (the sample size).

Before sampling we needed to define the word form on which to base the type counts. The alternative options are the raw word form or the lemma level. A lemma, also called a base form, is the form in which a word is listed in a dictionary. Word forms, on the other hand, are the 'actual manifestations' of a lemma. For example, tokens like 'run', 'runs', 'running' and 'ran' belong to the same lemma, viz. 'run', but are four distinct word forms. Most research in applied linguistics measures lexical richness on lemmas instead of word forms (e.g., Laufer 1991; Engber 1995), arguing that the inflected word forms indicate 'grammatical knowledge' rather than pure lexical knowledge (Vermeer 2000: 74). This might be a valid argument in studying the lexical use of children or L2 speakers, who have not fully acquired the grammar of the language. For adult native speakers, however, we may expect that we run the risk to confound lexical with grammatical knowledge by using the word forms as lexical units. Evert and Baroni (2005) conclude that analyses on lemmas and word forms give similar results in their study on word frequency distributions in the BNC. Gries (2006) observes in a quantitative study of the semantics of English ditransitives that the analyses on inflected verb forms and on lemmas yield the same results. He draws the conclusion that there are no urgent theoretical or practical reasons to opt for lemmas instead of word forms.

We used the word forms as they are distinguished in the CGN, including the part-of-speech tag. For example, a word form like *stap* ('step') can either be a verb (i.e. a form of the verb *stappen* ('to step')) or a noun (i.e. *een stap* ('a step')). Depending on the part-of-speech tag (or POS-tag) following the word form, the token is distinguished as either *stap/verb* or *stap/noun* (Van Eynde, Zavrel and Daelemans, 2000).

What is the optimal sample size (number of word tokens) for investigating the number of types (= TTR, given the fixed sample size) in our study? Neither in applied linguistics nor in mathematical linguistics, an 'established' token length exists for measuring lexical richness. Nevertheless, previous research provides global guidelines for the determination of a token length. Measuring on samples that are small, as is often done in applied linguistics (e.g. 35-50

tokens, Malvern et al., 2004), is not advisable, since this results in artificially inflated and instable TTRs. Tweedie and Baayen (1998) used text samples of 2000 tokens. These samples were drawn from full-sized novels, such as Lewis Carroll's Alice in Wonderland. Van Gijsel (2007) measured the TTR on a series of different token sizes, ranging from 150 to 1500 tokens by steps of 150, on samples from the CGN. In applying linear regression analyses, the impact of speaker variables turned out to become stable at the size of 1350 word tokens.[5] Applying this size, our smallest subcorpus is only large enough to create eight different samples of 1350 words. We decided to draw 10 samples from each of the other 23 subcorpora. That means that the resulting data file for further analysis contained 238 samples.

We drew random (without replacement) and non-random samples from the subcorpora. The non-random samples were drawn by starting at the beginning of the subcorpora, taking subsequent text parts of 1350 word tokens. We called these samples the context samples, as the word tokens occur in their original context (their original order). We did the same for the random samples, after having randomized however the order of all word tokens of the whole subcorpus.

All lexical measures were computed for each of the samples. The number of hapax legomena is the number of words appearing only once in the sample at hand. The number of types is the number of different words in a sample. Since the CGN is a tagged corpus, we could easily classify and count words as nouns, verbs or adjectives. Adverbially used adjectives are POS-tagged as adjectives. We made use of the list of most common words as constructed from the Spoken Dutch Corpus by Keune et al. (2005).

The outcomes of the six lexical measures are given in Table 5.1. For each measure the mean of all 238 samples is calculated, together with its standard deviation, plus the probability outcome of a test on the normality of the resulting distribution (Shapiro Wilk). A value lower than .05 indicates a significant deviance from the normal distribution. A distinction is made between the random and the context samples.

Table 5.1 shows clear differences in the counts of types and hapax legomena between the random and context samples. The differences are reducible to the sampling frame. The context samples contain coherent text parts, with continued topics and repeating words. The other four lexical measures have no substantial differences. Their outcomes are not dependent on the sampling frame. The standard deviations differ for all lexical measures, all higher outcomes occurring in the context samples. Selecting coherent text parts also implies that speaker or topic specific variation is included in the variance, leading to larger differences. Such differences are neutralized in the random samples. This conclusion is corroborated by the outcomes on the normality tests. The random samples of the lexical measures all produced normal distributions. The context samples produced skewed distributions for nouns and adjectives, indicating that

[5]See Van Gijsel (2007) for further details. This study was carried out on a larger part of the CGN corpus, also testing the effect of register.

Table 5.1: Mean number of counts, standard deviation, and normality assumption (Shapiro-Wilk Normality test; p value, $p < 0.05$ significant) for each of the six lexical measures. The left columns give the results for the random samples, the right columns the results for the context samples. (sd = standard deviation.)

	RANDOM SAMPLES			CONTEXT SAMPLES		
measures	counts	sd	normality	counts	sd	normality
hapaxes	345.4	21.6	0.6116	254.5	29.8	0.0929
types	497.6	19.8	0.5736	426.5	31.9	0.6586
mcws	822.5	27.8	0.1961	816.4	44.2	0.8555
nouns	130.2	14.6	0.7253	132.8	25.8	0.0041
adjectives	73.6	11.5	0.3705	75.3	18.9	0.0037
verbs	228.9	16.2	0.8630	230.5	20.0	0.6613

nouns and adjectives may have relatively higher frequencies in coherent text parts. The results point out that random samples have more stable properties, which is a compelling argument to base further analyses primarily on the random samples.

5.2 Results

We started the analysis of the data by applying a Principal Component Analysis (PCA). The aim was to obtain an overview of the global patterns in the multivariate distribution of the six lexical measures and the four social variables. How are the lexical measures interrelated and how do their relational patterns co-vary with the social variables? Both the 238 random and context samples yielded a factor solution with three outspoken principal components (with an eigenvalue higher than or near to 1). The factor solutions after rotation (varimax) appeared to be similar, the random samples returning a more outspoken structure. We will present the results for the random samples.

Figure 5.1 summarizes the results of the PCA on the random samples. The first three principal components each explain more than 10 per cent of the variance: respectively, 46.2%, 24.2%, and 16.1%. The corresponding eigenvalues were respectively 1.665, 1.205, and 0.983.[6] The three right panels visualize the loadings of the lexical measures on the three components or dimensions after varimax rotation. On all three components, types and hapax legomena are in each other's neighborhood, indicating that they are tightly related measures in our corpus. They have high loadings on PC1, in fact the highest loadings of

[6] In order to test the stability of the data, we selected 100 samples of each subcorpus (with replacement) instead of 10. The results were similar.

all variables, indicating that this dimension can be qualified as lexical richness. Verbs and adjectives have no relationship with richness, but nouns have. The most common words (MCWs) have a self-evident negative load on the first dimension.

Principal component 2 is marked by a high positive loading of the verbs, in combination with a negative loading of adjectives plus MCWs. Principal component 3 gives a clear contrast between verbs plus adjectives versus nouns. We hope to obtain better insight how these contrasts can be explained by applying regression analyses to the data.

The left panels of Figure 5.1 plot the 238 samples in the lexical space as spanned by the three principal components. We can search for sociolinguistic structure by visualizing the properties of the samples. In each panel we visualize the most outstanding effects on the horizontal principal component. As we can see in the upper left panel, PC1 captures gender differences: female speakers, mainly positioned on the left side of the panel, use MCWs, while male speakers, mainly positioned on the right, use more hapax legomena and a greater variety of word types.

Principal Component 2 captures the aspects of country origin. This variation is an effect of the difference in the use of adjectives and MCWs on the one, and verbs on the other side (loadings resp. -0.55, -0.45 and 0.66). The use of many adjectives and MCWs characterizes spontaneous speech from the Netherlands.

Principal Component 3, visualized on the horizontal axis of the lower left panel, captures age differences. While most speech fragments of young speakers are situated more to the right end of the panel, the speech fragments of the middle aged and old speakers are situated more to the left end of the panel. This variation seems an effect of the variety in the use of the different POS categories: old and middle aged speakers use more nouns, whereas young speakers use more verbs and adjectives (loadings resp.: -0.43, 0.51, and 0.71).

We need to unravel the impact of the sociolinguistic variables in more detail, including the way they may interact. The data were analyzed by applying linear modeling (GLM), the lexical measures being the criterion variables, with country, gender, education level and age as the predictors. We included all possible interactions. Such an analysis allows us to check whether our first interpretations on the basis of the factor analysis are correct and robust. We discuss the results for each lexical measure separately. We did not include all significant effects in our discussion, but were stricter by adding a criterion of effect size. The problem with the many data we have is that given sufficient samples almost all possible effects tend to reach the level of statistical significance (in our case, alpha = 0.05). Even fairly accidental effects may reach the level of significance given larger sample sizes or higher numbers of samples obtained from the corpus. We decided to separate the substantial effects from the more marginal and accidental ones by adding a criterion of effect size: we took only effects into account that passed the criterion of 10% explained variance

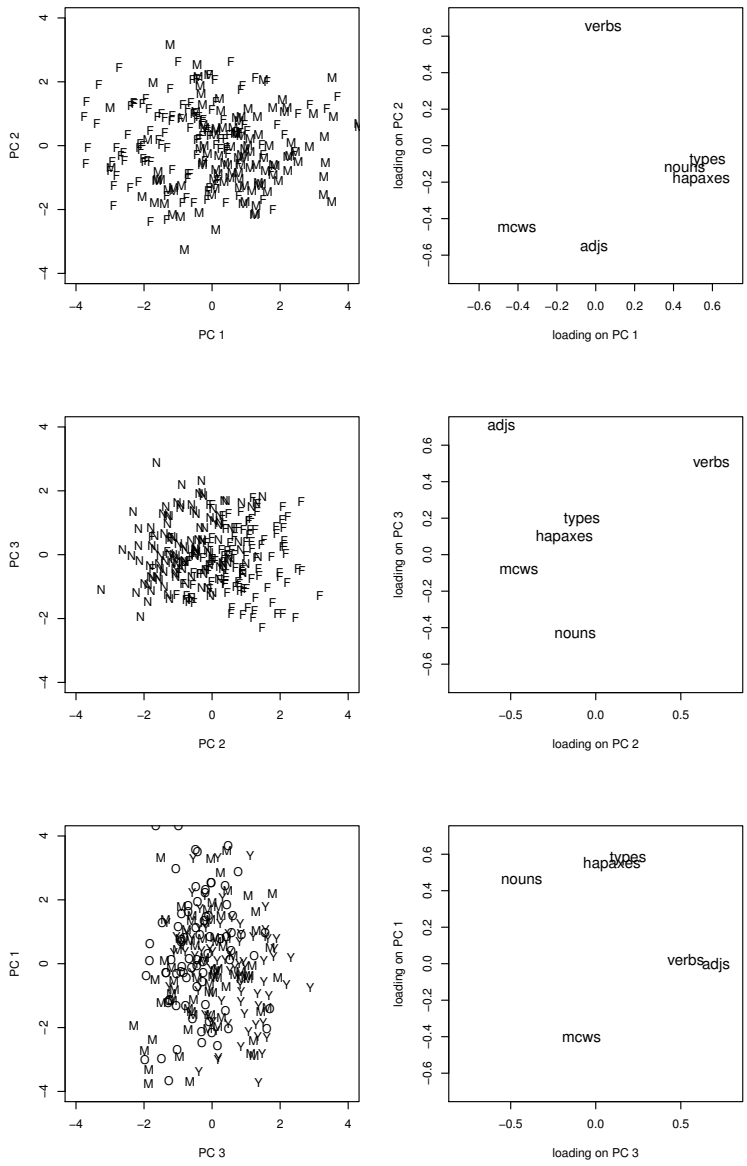


Figure 5.1: Principal Component Analysis of the 238 samples from 24 subcorpora of spontaneous Dutch speech. In the left panels the dimensions of the 238 samples are plotted. In the right panels the loadings of the lexicon measures on the 238 samples. (adjs = adjectives)

(criterion applied is the partial eta squared, η^2).

A GLM analysis on the number of hapax legomena occurring in the 10 samples of each of the 24 subcorpora revealed a main effect for gender: men used hapax legomena more frequently than women (resp. 354.6 versus 336.4 per sample; $F(1, 214) = 76.306$, $p = 0.000$, $\eta^2 = 0.263$). This outcome suggests a higher lexical creativity of male speakers. Furthermore, an effect for education level was found: higher educated speakers used more hapax legomena than lower educated speakers ($F(1, 214) = 38.601$, $p = 0.000$, $\eta^2 = 0.153$). This effect for education was modulated by an interaction effect of country by education level ($F(1, 214) = 33.359$, $p = 0.000$, $\eta^2 = 0.135$). It appeared that in the Netherlands there was no effect of education at all. In Flanders however, the result was substantial: higher educated speakers used a mean of 359.1 hapax legomena, and lower educated speakers 339.2.

Next, we fit a linear model to the number of types occurring in our 238 samples of spontaneous speech. The results were similar to those for the hapax legomena. This does not come as a surprise, since hapax legomena and types has a correlation of 0.93. However, the total amount of variation explained by the hapax legomena was higher. ($\eta^2 > 0.10$; see Table 5.2, adjusted R^2 resp. 0.482 and 0.431). For this reason, we prefer to work with the number of hapax legomena.

A linear model applied to the number of most common words occurring in the subcorpora revealed that most variation was explained by country: Dutch speakers used most common words more frequently than Flemish speakers (resp. 839.7 versus 805.0; $F(1, 214) = 203.638$, $p = 0.000$, $\eta^2 = 0.448$). In addition, there was a significant effect for gender: female speakers used MCWs more frequently than male speakers (resp. 828.9 versus 816.1; $F(1, 214) = 28.512$, $p = 0.000$, $\eta^2 = 0.118$).

Next, we applied a linear model analysis to each of the three categories of POS: nouns, adjectives and verbs. It turned out that these three word categories do not behave similarly. The results for nouns revealed significant and substantial effects for gender (resp. 134.4 versus 126.1; $F(1, 214) = 27.857$, $p = 0.000$, $\eta^2 = 0.115$) and age (young 126.8, middle aged 128.1, old 135.8; $F(2, 214) = 12.217$, $p = 0.000$, $\eta^2 = 0.102$). Men used nouns more frequently than women, and old speakers used nouns more frequently than both middle aged and young speakers (resp. $F(1, 142) = 15.197$, $p = 0.000$, $\eta^2 = 0.134$, and $F(1, 142) = 22.040$, $p = 0.000$, $\eta^2 = 0.097$)).

Most variation in the use of adjectives is explained by country ($F(1, 214) = 145.426$, $p = 0.000$, $\eta^2 = 0.405$). Dutch speakers used adjectives more frequently than Flemish speakers (mean: 79.6 versus 67.5 adjectives). The speaker's age too is an important predictor of the number of adjectives used (young 79.7, middle aged 71.6, old 69.3; $F(2, 214) = 39.815$, $p = 0.000$, $\eta^2 = 0.271$). Young speakers used adjectives more often than middle aged and old speakers (resp. $F(1, 144) = 45.586$, $p = 0.000$, η=.240, and $F(1, 142) = 69.965$, $p = 0.000$, $\eta^2 = 0.0330$).

Table 5.2: Strong effects (significant ($\alpha = 0.05$) and $\eta^2 > 0.10$) in the six lexical variables. In the upper panel the effects for the random samples are given. The lower panel presents the effects for the context samples. (adj. R^2 = adjusted R^2)

RANDOM SAMPLES					
	gender	country	age	education	adj. R^2
hapaxes	men	-	-	Flemish high	0.482
types	men	-	-	Flemish high	0.431
mcws	women	Netherlands	-	-	0.542
nouns	men	-	old	-	0.331
adjectives	-	Netherlands	young	-	0.532
verbs	women	Flanders	-	-	0.235

CONTEXT SAMPLES					
	gender	country	age	education	adj. R^2
hapaxes	-	-	-	-	0.163
types	-	-	-	-	0.173
mcws	-	Netherlands	-	-	0.251
nouns	men	-	-	-	0.278
adjectives	-	Netherlands	young w.r.t. old	-	0.295
verbs	-	-	-	-	0.224

The speaker's gender and country explained most variation for verbs. Women use verbs more frequently than men (resp. 233.5 versus 224.3; $F(1, 214) = 24.627$, $p = 0.000$, $\eta^2 = 0.103$), and Dutch speakers use more verbs than Flemish speakers (resp. 233.6 versus 224.3; $F(1, 214) = 25.123$, $p = 0.000$, $\eta^2 = 0.105$). The results are summarized in Table 5.2. We included the adjusted R^2 explained by each model to indicate the total effect size.

Table 5.2 gives the effects for both the random and context samples. The differences are as expected. The effects for the context samples are a subset of the effects of the random samples and the effects remaining have smaller effect sizes. Random sampling our subcorpora has more power for the same number of lexical elements samples and the effects found are more stable, resulting in higher amounts of explained variance.

The linear model analyses corroborate the outcomes of the principal component analysis. However, there are two noticeable differences, both concerning the sociolinguistic variation in verb frequency. First, in the PCA analysis the number of verbs occurring in the different samples loaded high on PC3, indicating an age effect. However, our linear regression model for verbs, revealed no effect for age at all ($F(2, 214) = 0.302$, $p = 0.740$, $\eta^2 = 0.003$). Second,

our linear model analysis of the verbs did reveal a highly significant effect for gender, which was not visible in the PC analysis.

In order to obtain a better understanding of our results for the variation in the different POSs, we take a closer look at our data by investigating the contribution of individual words. The effect of lexical measures can be the consequence of all the words involved in a measure, implying that there is a global effect affecting all lexical items. On the other hand, the impact of a lexical measure may be brought about by a specific word or subsets of words, the frequent words in particular. In that case, the effect has to be interpreted as word-bound, and not as an effect that is characteristic of all words included in a lexical measure. Inspection of the frequency distributions of the 20 most frequently used nouns revealed no specific nouns that can explain the overall variation we found. There were no effects for the 20 most frequent words when analyzed separately in a linear model analysis. It seems that men and old speakers use more nouns in general than the other speakers. The effect is global and not word-bound.

Next, we explored the 20 most frequent adjectives. We ran general linear models on these 20 adjectives to trace the effects of country and age. The words *lekker* ('nice', 'exquisite'), *mooi* ('beautiful'), *leuk* ('nice', 'fine'), *gewoon* ('ordinary', 'just'), *precies* ('exactly'), *erg* ('very'), *heel* ('very'), and *hele* ('very'), appeared to be words typical of Dutch speakers, and *gewoon* ('ordinary', 'just'), *echt* ('really'), and *leuk* ('nice'), turned out to mark young speakers. We reanalyzed our data for adjectives and excluded the counts for the markers in our model. Without these makers there was neither an effect for country, nor for age: these effects appeared to be word-bound.

We explored whether specific verbs in the set of 20 most frequent verbs mark country and gender, whereas others do not. Single verbs did not mark a gender difference. Women use more verbs than men in general. For country however, five verbs revealed significant effects with an η^2 larger than 0.1, namely *is* ('is'), *zijn* ('are'), *'s* ('s' < 'is'), *gaan* ('go' (plural)), and *hebt* ('have' (singular)). All these verbs were used more frequently in Flanders. The forms of the verb 'to be' revealed the strongest effects. The removal of only the counts for these three verbs from our total counts for verbs, resulted in a model without any effect for country. Apparently, the variation found between Flanders and the Netherlands is mainly an effect of the use of the 3rd singular, and plural word form of the verb 'to be'.

Finally, for the most common words, we found a word-bound effect for country. The words *ik* ('I'), *je* ('you'), and *uh* ('uh', a hesitation marker), were typical of Dutch speakers. Without these three words, the effect for country disappeared. The words *wel* ('well', 'indeed'), *ook* ('too', 'also'), *dan* ('then'), and *een* ('a') emerged as typically Dutch too. The words *dat* ('that'), and *is* ('is'), appeared as typical words of Flemish speakers. For three out of the 20 most frequent words, we found a significant effect for gender: the words *uh* ('uh', hesitation marker), *een* ('a'), and *ik* ('I') were typically used by women.

Table 5.3: Type of effect in four lexical measures: word-bound (word-specific) versus global.

	gender	country	age	education
nouns	global	-	global	-
adjectives	-	word-bound	word-bound	-
verbs	global	word-bound	-	-
mcws	global	word-bound	-	-

Removing these words from our data did not take out the effect for gender. Apparently, the effect of gender is global.

In Table 5.3 we give an overview of the results after having explored the 20 most frequent nouns, adjectives, verbs, and MCWs. Gender effects are global for all POSs, while variation between Flanders and the Netherlands is word-bound. The age effect for adjectives is clearly word-bound. For nouns, however, this effect is global.

5.3 Conclusion and discussion

The aim of the present study was to investigate the sociolinguistics patterns of global lexical variation in spontaneous speech and the social stratification of such lexical characteristics. Four social variables were investigated in a corpus of spontaneous spoken Dutch (CGN): country (Flanders vs. the Netherlands), gender, education level and age. We distinguished three types of global lexical measures: lexical diversity (hapax legomena and number of types), lexical communality (the frequency of use of the MCWs), and lexical density (the frequency of use of nouns, adjectives and verbs). To obtain a better understanding of the patterns of variation we found, we investigated additionally whether the outcomes on the lexical communality and density measures were determined by specific words only (word-bound patterns) or that all words were part of the pattern (global or general patterns, as they apply to the whole class or group of words involved).

We successfully used principal component analysis to explore globally if and how the lexical measures were related, both to each other and to the four social variables. The PCA proved to be a valuable technique to obtain a general overview of the different sources of variation in our data. Next, we applied linear model analysis (GLM) on each of our six lexical measures, to test the impact of our four social variables. The social patterns pointed out by the principal component analysis turned out to be indicative of the outcomes of the separate lexical measures. However, the linear models gave more detailed information on the separate social variables, their interactions, and their effect sizes. All social variables had an impact on one of the lexical measures as a main effect. We

only found one interaction effect, country by education. The effect sizes were considerable (more then 0.10) and for the random samples, the adjusted R^2 ranged between 0.235 and 0.542. There is substantial lexical variation that is tightly embedded in patterns of social variation. Below, we discuss the effects of the social variables as we have observed them in the random samples.

Gender came out as the most outspoken and present variable. Men used more hapax legomena, types and nouns, whereas female speakers use more verbs and MCWs. Moreover, the differences that occurred appeared to be global: within the 20 most frequent words there was no subset of words by which most variation was explained. The general character of gender may explain why gender systematically emerged as an important factor of language variation in previous research (Newman et al., 2008; Härnqvist et al., 2003).

The finding that men used more hapax legomena, types, and nouns, is in line with previous research on gender differences. In register research, Biber (1988) applied dimension reduction techniques, and found that the most important dimension determining register variation is the 'involved versus informational' continuum. Linguistic features indexing a more 'informational style' include, among others, a high Type-Token Ratio, a high rate of nouns, longer words, adjectives, and prepositions, whereas features as second person pronouns, present-tense verbs, diminutive affixes, that-deletions and contractions belong to an 'involved' register. Biber, Conrad and Reppen (1998) applied these findings to a corpus of personal letters of men and women from the 17^{th} through the 20^{th} century. Despite the fairly small number of letters, Biber found a higher index of involved language in personal letters of women in both the 17^{th} and the 20^{th} century.

Biber's findings are in agreement with the results obtained by Newman et al. (2008), who explored gender variation in language use on a large language corpus of, mainly written, data. They found that female language included more pronouns and social words, a wide variety of other psychological references, and more verbs. Male speech, on the other hand, was more likely to include long words, numbers, articles, and prepositions. Men discussed current concerns more frequently and swore more often. They concluded, 'Female language emphasized psychological processes, social processes, and verbs. Male language emphasized current concerns.' Another study in line with our results is Rayson et al. (1997) who used the spoken data of the British National Corpus. They conclude that men used more nouns, while women used more verbs. The more frequent use of most common words by women is in line with previous research by Keune et al. (2005). As can be seen from the loadings on PC1 in our principal component analysis, the number of hapax legomena, types, and nouns that are used contrasts with the number of MCWs used. The list of MCWs comprises mostly function words (for instance pronouns, articles, prepositions, conjunctions, and auxiliary verbs), which are used at much higher rates in natural conversations than in written language, especially by women (Newman et al., 2008).

The recurrent and stable character of the gender effect seems to be grounded on a different style register what can be characterized as Biber did (1998), by 'informational' versus 'involved' language use. The best term to qualify such a distinction seems to be the concept of style, although we need to understand better how such a style distinction corresponds to language production processes that exploit the lexical resources differently.

In addition to the speaker's gender, country is an important predictor of lexical richness. We found that Dutch speakers used more MCWs and adjectives, while Flemish speakers used more verbs. In contrast to our findings for gender, all effects for country appeared to be word-bound instead of global. The adjectives that appeared as typically Dutch contained particular intensifiers, and specific words that can be used as interjections and discourse markers. This suggests that the differences between the speech communities can be traced back to divergent lexical choices in expressing specific concepts.

Our finding that the effect for country is word-bound is supported by research from, for instance, Geeraerts et al. (1999) who found that in the Netherlands and Flanders different words are used to express similar objects. The results meet our expectation that there are no global lexical differences between the two speech communities.

The most striking outcome for the variable of education is the absence of an overall main effect in all six lexical measures. We are reluctant in predicting a difference, as we investigate informal speech. Highly educated speakers do not fully exploit all their lexical resources in private, non-professional spontaneous speech. Private registers apparently show less difference in lexical richness and lexical style than audience oriented or public speech.

We only found variation in the level of education in the use of the number of hapax legomena and types within Flanders, with higher scores for higher educated speakers in comparison to lower educated speakers. This may be explained by the problematical standardization process in Flanders, which has resulted in a 'linguistic insecurity' for the Flemish speakers (Taeldeman, 1992; Geeraerts, 2001, 2003). This insecurity mainly appears in situations typically requiring the standard language. Due to this insecurity, Flemish speakers use a more formal speech level than Dutch speakers, for whom this insecurity is not present. As already mentioned above, this formality may result in the use of more specific and complex words. This effect is strengthened by the comparison of highly educated Dutch and Flemish speakers: the lexical richness of Flemish highly educated speakers was significantly higher than that of the Dutch highly educated speakers.

The variable age showed two effects. First, a global age effect for nouns was revealed. The fact that the old speakers used nouns more frequently than middle aged and young speakers, may suggest that they use the language in a less involved way. If this were the case, it should have been a more pervasive effect in other lexical measures as well. However, we did not obtain the expected higher lexical richness for older speakers. We expect old speakers to use more nouns,

since they have the advantage of having expanded their lexical knowledge over a longer time span, including all lexical innovations and names for new products and tools. Nevertheless, this effect needs to be investigated further, as it turns out to be a global effect. The age effect that emerged for adjectives appeared to be typically word-bound. Young speakers used the words *gewoon* ('ordinary') and *echt* ('real') as popular interjections and discourse markers, for instance in collocations as *gewoon leuk* ('just fun') and *ja, echt* ('yes, really'). Our results confirm the conclusion by Rayson et al. (1997), who found that young speakers make intensive use of some specific adjectives.

We mentioned earlier that the selection of informal spontaneous speech obtained in private, non-professional telephone and face-to-face conversations may imply a reduction of differences between speakers with varying educational backgrounds. Spontaneous speech spoken in informal situations converges at the global lexical level, the level we operationalized with our lexical measures. That does not exclude the frequent occurrence of words in the corpus we investigated (CGN) that are more typical of particular social classes, social groups or educational groups. Word-bound effects may trigger variation in global lexical measures, but many word-bound effects will remain unnoticed by our lexical measures. Patterns of variation for individual lexical items may be interesting from a sociolinguistic point of view, as they will often be embedded in social structures, but they were not the target of the present study. The same strong social embedding and stratification was found in sociolinguistic studies for many phonological and morpho-syntactic variables, where the most outspoken differences were found in spontaneous, informal speech. The global lexical measures seem to stratify between social groups better when they are studied in formal contexts. There is no data available yet to confirm this interpretation. What we do know is that the six lexical measures yield large differences when different registers are being compared (Van Gijsel et al., 2005, 2006).

The sociolinguistic point of view brings us back to the gender effect that turned out to be systematic and global. A recurrent finding in sociolinguistics is that women are more norm or prestige oriented in their choices of variants of linguistic variables. Such an orientation cannot be the automatic outcome of a more involved, partner oriented way of speaking as compared to men who, in return, are more informationally driven or led by motives reducible to their wish to make an impression on their audience (impression management). The gender difference must be explained on a deeper level of communicative behavior, including perhaps a more coherent and careful exploitation of lexical resources and including more prestige oriented phonological and morpho-syntactic choices by women, having the consequence that women's speech is more transparent, coherent and accessible.

At the end of this final section, we return to a few methodological aspects of our study. We motivated the use of samples drawn from the relevant subcorpora by referring to the text-length dependency of the measures of lexical richness and their liability to topicality. The subcorpora we created by stratifying

the corpus of spontaneous Dutch speech into 24 subcorpora that substantially varied in size (from 10,966 words to 724,811 words). Working with these subcorpora varying in size would have caused unreliable results, probably in other lexical measures as well (Tweedie and Baayen, 1998). We therefore decided to take samples from these subcorpora, each containing the same number of words tokens. We drew random samples (drawing words randomly, without replacement, from the whole subcorpus) and context samples (drawing connected text parts form the subcorpus). Random samples gave the analyses much more statistical power (the lexical measures having much lower standard errors, plus normally distributed) and they appeared to have superior qualities for tracing sociolinguistic patterns. We can be less sure about the sample size we took (1350) and the number of samples (10) per subcorpus. The sample size was determined by calculations on the power of sample sizes in the subcorpora of the CGN. The size of 1350 words was large enough to reveal sociolinguistic effects, but more investigation is needed to determine optimal sample sizes for more general measures in corpus studies. The number of samples (10) was determined by the size of the smallest subcorpus. Additional studies, including simulations studies, are required to get a better grip on optimal sampling frames in corpus studies.

References

Argamon, S., M. Koppel, J. Fine and A. Shimony, 2003. Gender, genre, and writing style in formal written texts. *Text*, 23 (3)

Arnaud, P. J. L., 1984. The lexical richness of L2 written productions and the validity of vocabulary tests. In T. Culhane, C. Klein-Bradley and D. Stevenson, eds., *Practice and Problems in Language Testing. Papers from the International Symposium on Language Testing.* Colchester: University of Essex, 14–28

Baayen, R. H., 1994. Derivational productivity and text typology. *Journal of Quantitative Linguistics*, 1: 16–34

Baayen, R. H., 1996. The effect of lexical specialization on the growth curve of the vocabulary. *Computational Linguistics*, 22: 455–480

Baayen, R. H., 2001. *Word Frequency Distributions.* Kluwer Academic Publishers, Dordrecht

Baayen, R. H., H. van Halteren, A. Neijt and F. J. Tweedie, 2002. An experiment in authorship attribution. In *Proceedings of JADT 2002.* Université de Rennes, St. Malo, 29–37

Baayen, R. H., H. van Halteren and F. Tweedie, 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11: 121–131

Biber, D., 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge

Biber, D., 1995. *Dimensions of Register Variation*. Cambridge University Press, Cambridge

Biber, D. and S. Conrad, 2009. *Register, Genre, and Style*. Cambridge University Press, Cambridge

Biber, D., S. Conrad and R. Reppen, 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, Cambridge

Burrows, J. F., 1992a. Computers and the study of literature. In C. S. Butler, ed., *Computers and Written Texts*. Blackwell, Oxford, 167–204

Burrows, J. F., 1992b. Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7: 91–109

Burrows, J. F., 1993a. Noisy signals? Or signals in the noise? In *ACH-ALLC Conference Abstracts*. Georgetown, 21–23

Burrows, J. F., 1993b. Tiptoeing into the infinite: Testing for evidence of national differences in the language of English narrative. In S. Hockey and N. Ide, eds., *Research in Humanities Computing '92*. Oxford University Press, London

Coates, J., ed., 1998. *Language and gender: A Reader*. Blackwell, Oxford

Engber, C. A., 1995. The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4 (2): 139–155

Evert, S. and M. Baroni, 2005. Testing the extrapolation quality of word frequency models. In P. Danielson and M. Wagenmakers, eds., *Proceedings of Corpus Linguistics 2005*. Birmingham

Geeraerts, D., 2001. Een zondagspak? Het Nederlands in Vlaanderen: gedrag, beleid, attitudes. *Ons Erfdeel*, 44: 337–344

Geeraerts, D., 2003. Rationalisme en nationalisme in de Vlaamse taalpolitiek. In J. De Caluwe, D. Geeraerts, S. Kroon, V. Mamadouh, R. Soutaert, L. Top and T. Vallen, eds., *Taalvariatie en Taalbeleid. Bijdragen aan het Taalbeleid in Nederland en Vlaanderen*. Garant, Antwerpen and Apeldoorn

Geeraerts, D., S. Grondelaers and D. Speelman, 1999. *Convergentie en Divergentie in de Nederlandse Woordenschat. Een Onderzoek naar Kleding- en Voetbaltermen*. Meertens Instituut, Amsterdam

Gries, S. T., 2006. Exploring variability within and between corpora: Some methodological considerations. *Corpora*, 1 (2): 109–151

Grondelaers, S. and R. van Hout, 2011. The standard language situation in the Low Countries: Top-down and bottom-up variations on a diaglossic theme. *Journal of Germanic Linguistics*, 23 (3): 199–243

Härnqvist, K., U. Christianson, D. Ridings and J.-G. Tingsell, 2003. Vocabulary in interviews as related to respondent characteristics. *Computers and the Humanities*, 37: 179–204

Heylighen, F. and J.-M. Dewaele, 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7 (3): 293–340

Holmes, D. I., 1994. Authorship attribution. *Computers and the Humanities*, 28 (2): 87–106

Jaspaert, K., 1986. *Statuut en Structuur van Standaardtalig Vlaanderen*. Universitaire pers, Leuven

Johansson, V., 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. In *WorkingPapers*, volume 53. Lund University, Dept. of Linguistics and Phonetics, 61–79

Keune, K., M. Ernestus, R. van Hout and R. H. Baayen, 2005. Social, geographical, and register variation in Dutch: From written 'mogelijk' to spoken 'mok'. *Corpus Linguistics and Linguistic Theory*, 1: 183–223

Keune, K., R. van Hout and R. H. Baayen, 2006. Socio-geographic variation in morphological productivity in spoken dutch: A comparison of statistical techniques. In J.-M. Viprey, ed., *Actes des 8es journées internationales d'analyse statistique des données textuelles*, volume 2. Presses Universitaires de Franche-Comté, 571–580

Laufer, B., 1991. The development of L2 lexis in the expression of the advanced language learner. *Modern Language Journal*, 75 (4): 440–448

Laufer, B. and P. Nation, 1995. Vocabularly size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16 (3): 307–322

Malvern, D. and B. Richards, 2002. Investigating accomodation in language proficiency in interviews using a new measure of lexical diversity. *Language Testing*, 19: 85–104

Malvern, D. D., B. J. Richards, N. Chipere and P. Durán, 2004. *Lexical Diversity and Language Development*. Palgrave Macmillan

Newman, M. L., C. J. Groom, L. D. Handelman and J. W. Pennebaker, 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45: 211–236

Oostdijk, N. H. J., 2002. The Design of the Spoken Dutch Corpus. In P. Peters, P. Collins and A. Smith, eds., *New Frontiers of Corpus Research*. Rodopi, Amsterdam, 105–112

Plag, I., C. Dalton-Puffer and R. H. Baayen, 1999. Productivity and register. *Journal of English Language and Linguistics*, 3: 209–228

Rayson, P., G. Leech and M. Hodges, 1997. Social differentiation in the use of English vocabularly: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2 (1): 133–152

Richards, B., 1987. Type/Token ratios: What do they really tell us? *Journal of Child Language*, 14: 201–209

Smith, J. A. and C. Kelly, 2002. Stylistic constancy and change across literary corpora: Using measures of lexical richness to dat works. *Computers and the Humanities*, 364: 411–430

Taeldeman, J., 1992. Welk Nederlands voor Vlamingen? *Nederlands Van Nu*, 40: 33–52

Tweedie, F. J. and R. H. Baayen, 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32: 323–352

Van den Toorn, M. C., W. Pijnenburg, J. A. van Leuvensteijn and J. M. van der Horst, 1997. *Geschiedenis van de Nederlandse Taal*. Amsterdam University Press, Amsterdam

Van Eynde, F., J. Zavrel and W. Daelemans, 2000. Part of speech tagging and lemmatisation for the Spoken Dutch Corpus. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Athens, 1427–1434

Van Gijsel, S., 2007. Sociovariation in Lexical Richness. A Quantitative Corpus Linguistic Analysis. Ph.D. thesis, Katholieke Universiteit Leuven

Van Gijsel, S., D. Speelman and D. Geeraerts, 2005. A variationist, corpus linguistic analysis of lexical richness. In *Corpus Linguistics 2005*. Birmingham

Van Gijsel, S., D. Speelman and D. Geeraerts, 2006. Locating lexical richness: A corpus linguistic, sociovariational analysis. In *Proceedings of JADT 2006*. Besancon: Université de France-Comte, 961–971

Van Hout, R. and A. Vermeer, 2007. Comparing measures of lexical richness. In H. Daller, J. Milton and J. Treffers-Daller, eds., *Modeling and Assessing Vocabulary Knowledge*. Cambridge University Press, Cambridge

Vandenbussche, W., 2010. Standardisation through the media. The case of Dutch in Flanders. In P. Gilles, J. Scharloth and E. Ziegler, eds., *Variatio Delectat. Empirische Evidenzen und theoretische Passungen sprachlicher Variation*. Peter Lang, Frankfurt am Main, 309–322

Vermeer, A., 2000. Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17 (1): 65–83

Wilcox, R. R., 2010. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*. Springer, 2nd edition

CHAPTER 6

Conclusion and discussion

6.1 Overview

The main aim of this thesis was to expand our knowledge on the role of register and sociolinguistic factors (country, gender, age, and education level) in shaping the way lexical characteristics vary in both written and spoken Dutch. Register relates to language *use*, the sociolinguistic factors to the characteristics of the language *user*. Corpus linguistics and stylistics concentrate on the effects of register on global text characteristics. The emphasis in variationist studies in sociolinguistics is put on the impact of social factors on specific linguistic variables that are being defined most of the time on the phonological or morpho-syntactic level. In this dissertation we have tried to combine the fields of sociolinguistics and corpus linguistics in studying general or global patterns of lexical variation in two large corpora.

We used newspaper articles from the CONDIV corpus (Grondelaers et al., 2000), and speech from the Spoken Dutch Corpus (CGN) (Oostdijk et al., 2002). The CONDIV corpus contains approximately 17.6 million words from three Dutch and four Flemish newspapers: For each country it contains articles from a quality newspaper, a tabloid, and a regional newspaper (for Flanders two), which made it possible for us to distinguish between country and register.

The Spoken Dutch Corpus (CGN) contains approximately 8.9 million words from speech fragments of Dutch and Flemish adults. Speech from various speech registers is included, the most interesting from a sociolinguistic perspective being: Private speech (unscripted conversations and telephone dialogues: 4.7 million words), and public speech (3.4 million words). The public speech part can be split up into dialogues (for instance debates, meeting, and interviews: 2.3 million words) and monologues (news, reportages, and commentaries (all broadcast), reviews, ceremonious speeches, and lectures: 1.1 million words). The CGN contains further speaker information.

In Chapter 2 we explored variation of register and country in the frequency of use for 80 Dutch words ending in the suffix *-lijk* in written newspaper Dutch, and variation of register, country, gender, education level, and age in the frequency of use and the degree of acoustic reduction of 32 Dutch words ending in the suffix *-lijk*.

This suffix is hardly productive anymore and can be qualified as lexicalized. For instance, *natuurlijk*, literally 'nature-like', usually means 'of course'. Because of the loss of compositionality and information density of these words, they are liable to strong acoustic reduction in their pronunciation: *natuurlijk* is frequently produced as *ntuuk* and even *tuuk*. A benchmark was created to test our hypothesis that variation in the use of one lexical category (the set of *-lijk* words) is also present in other characteristics of the lexicon. This benchmark was provided by the covariance structure among the 80 most common words (highest frequency words, MCWs), which tap into the syntactic habits of speakers (Burrows, 1992, 1993). We showed the advantages of statistical models using analysis of variance and covariance of lexical frequencies in factorially contrasted subcorpora in comparison to Principal Components Analysis. With these models it becomes possible to include many predictors simultaneously, to calculate the significance of the predictors, the possible interactions between predictors, and the effects of specific words. In written Dutch, we found similar patterns for register and sociolinguistic variation for both lexical categories (*-lijk* words and MCWs). In spoken Dutch we observed marked sociolinguistic differences. Speakers with a higher education level used words ending in the suffix *-lijk* more frequently, and women used MCWs more frequently. This result suggests that both lexical categories may tap into independent sources of variation.

It is worth noting that the frequency of occurrence of the individual words in *-lijk* and MCWs varied among all predictors in both written and spoken Dutch. For instance, in spoken Dutch the words *tamelijk* and *ongelooflijk* ('somewhat', 'unbelievable') appeared as more typical for men, while the words *vriendelijk* and *lelijk* ('friendly' and 'ugly') appeared as more typical for women. This word-specific behavior indicates that a lexical category does not constitute a fully coherent, structured set, but that words keep having their own autonomous lexical distribution. These distributions are, among others, influenced by register (use), and sociolinguistic factors (users).

Only 24 of the 32 different words in *-lijk* occurred sufficiently often in the corpus to take them in consideration for further study, and only 14 of these words in *-lijk* were evidencing reduction. The reduction was more prominent for men compared to women. In Flanders, highly educated speakers used fewer reduced forms than non-highly educated speakers. For country there was no global pattern related to the degree of reduction. There were word-specific reduction patterns however. The degree of reduction was co-determined by two linguistic factors, namely, the word's position in the sentence (final or non final), and the extent to which the word was predictable from its context. In

sentence final position words revealed little reduction, and words with a high predictability on the basis of the preceding word revealed more reduction. Interestingly, these two linguistic variables did not interact with the regional and social variables. This suggests that there are fundamental linguistic principles that operate in the same way across different registers, countries, and social characteristics. Our results show the importance of taking register, country, and social information into account in the explanation of acoustic reduction. We expect that in the future more high frequency words ending in *-lijk* will be liable to become effectively morphologically unanalyzed and therefore lexicalized. Our expectation is strengthened by the finding of Pluymaekers et al. (2005), that younger speakers tend to reduce words in *-lijk* to a greater extent than older speakers.

In Chapter 3, we investigated regional and social patterns of variation occurring in derivational potential productivity in Dutch spontaneous speech, and investigated what statistical model is most appropriate to analyze our data. For 72 affixes we extracted the derivational hapax legomena from the private and public speech subcorpora (approximately 8 million words) of the Corpus of Spoken Dutch. We analyzed the distribution of the hapax legomena over the 24 subcorpora, which were defined by the speaker's country (the Netherlands or Flanders), gender, education level (high versus non high), and age (young, mid or old). Due to the highly varying sizes of the subcorpora and the large number of cells with zero counts (the consequence of the sparseness of hapax legomena for several of the affixes), we were challenged to find a model that has a good fit to the data. We compared three different statistical models, an ordinary least squares linear model with a transformed proportion of hapax legomena in the subcorpus as the dependent variable, a linear mixed effects model with affix as random effect and again the transformed proportions as the dependent variable, and a generalized linear model with a binomial link function, considering the hapax legomena as successes and all remaining words as failures. This last model, the GLM, gave an excellent fit to our data. It outperformed the other two models on two points. First, inspection of the residuals showed us that it was more successful in predicting zero counts. This is due to the fact that this model is not constrained by the normality assumption that governs the distribution of residuals in ordinary least squares regression. Second, the Zipfian nature of affix productivity makes it impossible to treat affix as a random effect in a linear mixed effects model that implicitly assumes a random effect to follow a normal distribution with mean zero and unknown variance. For all three models we found significant effects pointing in the same direction for gender, education level, age, and affix, showing that highly educated old men revealed the greatest overall affix productivity. Furthermore, all three models revealed that there is no global regional and social variation pattern for all affixes. Many affix-specific differences occurred instead.

In Chapter 4 we investigated the effects of register and sociolinguistic factors on derivational and lexical variation in both written and spoken Dutch. We

made use of newspapers articles from the CONDIV corpus and speech from the Spoken Dutch Corpus. We created subcorpora distinguishing between register and country (the Netherlands versus Flanders) for written Dutch, additionally including gender, education level, and age for spoken Dutch. For both the written and the spoken corpus, we selected all derivational and lexical hapax legomena and assigned them to the subcorpus they belonged to. By counting the number of hapax legomena in a particular subcorpus, divided by the total number of words in the same subcorpus, we calculated the potential productivity or growth rate of that subcorpus. We hypothesized that within registers derivational productivity mirrors lexical productivity, which indeed turned out to be the case. Inspection of global lexical and of global derivational patterns in the register subcorpora of written Dutch (quality, national/tabloid, and regional newspapers) and spoken Dutch (formal monologue – public speech, dialogue – public speech, dialogue – private speech) demonstrated a clear and robust correspondence between derivational and lexical productivity. We showed that lexical productivity is for a very small part only the outcome of the presence of derivationally formed hapax legomena. Many other word formation processes are involved.

Lexical productivity turned out to be higher for written than for spoken Dutch. This is in line with previous research (cf. Biber and Conrad, 2009; Plag et al., 1999). We also found higher productivity in the more formal registers within written and spoken language. Derivational productivity, however, was not higher in written than in spoken Dutch. Even the removal of the diminutive affixes, which are highly frequent in spoken Dutch, and which came out as deviant members of the class of affixes, did not result in higher productivity scores for written Dutch.

The varying productivity of individual affixes across written and spoken Dutch could not adequately explain the relatively high overall affix productivity in spoken Dutch. An appealing explanation is that in spontaneous speech, speakers make a more active use of the productive properties of affixes to coin new words, because of time pressure in speaking where there is hardly time to consider and reflect about word choice, while in written Dutch writers may use well-established derivational words forms more frequently, because they have more time to consider lexical choices.

The highly frequent diminutives, with a productivity pattern that diverged from that of the other affixes, were highly influential in our analyses, and established an independ pattern all by themselves. This shows us the importance of closely inspecting the role of all affixes separately, before drawing general conclusions.

As in previous research (Biber and Conrad, 2009), register emerged as an important predictor of the degree of productivity: The more formal a register, the higher its degree of productivity. For country we found no global derivational and lexical differences, however we did find word-specific differences. This is in agreement with Keune et al. (submitted) who found that all lexical

differences between countries were word-specific.

Each of the social factors emerged as influential: Old men with a high education level used derivational and lexical items most productively. The relatively high productivity of male speakers is in agreement with previous research (Härnqvist et al., 2003; Van Gijsel, 2007). As in research by Keune et al. (submitted) all effects for gender turned out to be general and global. The higher productivity for highly educated speaker matches the higher productivity in quality newspapers (aimed at a higher educated readership). The degree of productivity increases with the speaker's age, but only when the speaker is highly educated. This shows that skills necessary to create new words and word forms develop over longer time span. Since private dialogues commonly have an informal and involved style, it is not surprising that the effect for age is absent in this speech style. It explains at the same time why Härnqvist et al. (2003), and Keune et al. (2006) found effects for age, while Keune et al. (submitted) who explored only speech from private dialogues, did not trace effects for education level.

In Chapter 5 we investigated the sociolinguistic patterns of global lexical variation in spontaneous speech from the Spoken Dutch Corpus. As in the previous chapter we investigated the variables of country, gender, education level, and age. We applied three types of global lexical measures: Lexical diversity (measured by the number of types and number of hapax legomena), lexical density (measured by counting the relative number of nouns, adjectives, and verbs), and lexical communality (measured by counting the frequency of occurrence of the most common words).

We divided the spontaneous speech corpus into 24 subcorpora according to the regional and social variables. To exclude effects of text-length dependency in the lexical diversity measures, we decided to work with text samples containing the same number of word tokens. We drew 10 random and 10 context samples, each consisting of 1350 words. The random samples, not liable to effects of topicality, gave the analyses much more statistical power, and had superior qualities for tracing sociolinguistic patterns.

Principal Component Analysis (PCA) proved to be a valuable technique for exploring the data. We used General Linear Models to test the significance and effect size of our four regional and social variables and to reveal possible interactions between these variables. We found main effects for all four variables, and one additional interaction. Gender proved to be the most outspoken explanatory variable. Men used more hapax legomena, more types, and more nouns, while female speakers used more verbs and used the most common words more intensively. These findings are in agreement with findings reported in previous research (Newman et al., 2008; Härnqvist et al., 2003; Rayson et al., 1997).

Further inspection of the data made clear that the effect for gender was global, as it could not be explained or related to specific lexical or word-bound effects only. This connects to the systematic gender effect found in sociolinguistic studies on language variation and change (Coates, 1998), where women

are ahead of men in innovation. Men seem to use a different speech style than women that can be characterized by the distinction of 'informational' versus 'involved' language use (Biber et al., 1998).

Country had a clear effect as well. Dutch speakers used more MCWs and adjectives, while Flemish speakers used more verbs. Considered jointly, all effects for country appeared to be word-bound, and can probably be traced back to divergent lexical choices or patterns in expressing specific concepts.

Concerning age, we found a global effect for the number of nouns, and a word specific effect for the number of adjectives used. Old speakers used more nouns than middle-aged and younger speakers. This may possibly be an effect of expanding lexical knowledge during the lifetime. Further research is needed to confirm this interpretation. The finding that young speakers used adjectives most frequently was typically a word-bound effect. It was caused by the more extensive use of specific adjectives, which were being used as popular interjections and as discourse markers by young speakers.

6.2 Register (use) and sociolinguistic (user) effects

6.2.1 Register

The distinction between written and spoken Dutch registers emerged as a substantial predictor of patterns of variation in the lexicon. In the most formal registers of spoken and written Dutch, we found the highest derivational and lexical productivity. We also found a higher global lexical productivity in written than in spoken Dutch. These findings are consistent with the 'informational versus involved' dimension that Biber and Conrad (2009) identified to contrast written and spoken language. However, for derivational productivity we did not find a substantially higher productivity in written Dutch. The productivity of the individual affixes highly varied across written and spoken Dutch: Some affixes were typically used in written Dutch, while other affixes were most characteristic for spoken Dutch. A possible explanation for the high productivity of some affixes in spontaneous speech is that speakers make more active use of the productive properties of affixes to coin new words. They have not much time to reflect on word choice and to reconsider what words can be produced best, and therefore tend to use productive affixes to facilitate lexical decisions. For written Dutch we additionally explored variation in the frequency of use of the suffix *-lijk* and the use of MCWs across different newspapers. Just as for derivational and lexical productivity, *-lijk* was most frequently used in the quality newspapers.

In the Netherlands there were no significant differences in the use of MCWs between the newspapers, and their use was comparable to their use in the Flemish regional newspaper. MCWs were used somewhat more often in the Flemish

tabloid, the most noticable difference is however the lower frequency of MCWs in the Flemish quality newspaper. This suggests that the quality newspaper in Flanders is stronger oriented towards a more dense style incorporating a larger range of lexical variation, perhaps supported by a tendency to avoid the communality of daily words.

6.2.2 Country

Our most important finding with respect to differences between the Netherlands and Flanders is that variation patterns are primarily word-bound. Words ending in *-lijk* turned out to be typically Dutch, and overall these words are more often pronounced in reduced form in the Netherlands. However, within in the set of words ending in *-lijk*, word-specific differences prevail. We even find word pairs that are (near) synonyms of which one word is typical of the Netherlands, and the other one of Flanders. The higher degree of reduction of words ending in *-lijk* in the Netherlands is word-bound as well, and the outcome of only four of the fourteen words we analyzed. The effects we found for derivational productivity were affix-bound. Inspection of the number of adjectives, verbs and MCWs in the speech samples we created for our study on global lexical variation (Chapter 5), strengthen our conclusion that differences between Flanders and the Netherlands are mainly word-bound. Verbs were used more frequently in Flanders, while adjectives and MCWs were used most often in the Netherlands. However, the removal of only a few words from these categories, made these effects disappear. The differences between the Netherlands and Flanders can probably be traced back to divergent lexical choices in expressing specific concepts. This is in agreement with previous research in which it is stated that except for a number of specific and highly frequent characteristics of 'Tussentaal' (see also Chapter 5), Dutch as spoken in Flanders is not at all that different from Dutch as spoken in the Netherlands (cf. Grondelaers and Van Hout, 2011).

6.2.3 Gender

Gender emerged, together with register, as the most important predictor of patterns of lexical variation. In contrast to our findings for country, lexical patterns for gender were primarily global: The removal of words which seemed most typical of men or women did not change our results. This connects to the systematic gender effect found in sociolinguistic studies (Coates, 1998). A high derivational and lexical productivity, a high Type-Token Ratio, and a high proportion of nouns, all characteristics of a more 'informational' speech style, characterized men's speech (Biber and Conrad, 2009; Rayson et al., 1997). A high proportion of verbs, and MCWs (mainly function words), typical of a less informational, more 'involved' speech style characterized women's speech. These findings are in agreement with previous findings by Newman et al. (2008), and Härnqvist et al. (2003). As for register, there were many word and affix-specific

patterns in the lexical characteristics we explored. A closer examination of the individual words and affixes that appeared as more typical of men or more typical of women confirmed our conclusion that men use more 'informational' lexical items, and that women more often express themselves in an 'involved' way of speaking. How can we further define the different speech styles of men and women? It is clear that the distinction between male and female speech is not a result of the use of individual words. The different speech styles of men and women seem to be mainly a social-cultural effect. This would mean that women use more involved speech than men because involved speech is considered as more appropriate for women in our society. On the other hand one may speculate that a more involved speech style for women is biologically determined as well, since across time and cultures, women are more intensively charged with the task to raise children. More interdisciplinary research is required to substantiate such a claim.

6.2.4 Education

The speaker's level of education was a predictor as well of the degree of productivity in speech and of the frequency of use of *-lijk* words. Apparently, for producing new words or exhausting lexical resources, skills are needed which are better available to highly educated than to non-highly educated speakers. These results are in line with research of Härnqvist et al. (2003), who found higher lexical richness outcomes for highly educated speakers. The affix-specific variation patterns only explained very little variation in relation to education, and were therefore not further considered. In our samples of private spontaneous speech (Chapter 5), the speaker's education level was not a relevant factor in explaining lexical variation in the Netherlands. This does not come as a surprise as the productive, lexical richer use of language needs to be triggered. As private spontaneous speech is generally more 'involved', it is likely that highly educated speakers do not exploit their full capacities in their informal speech register to use more infrequent and new words. In Flanders, we did find a higher lexical diversity in text samples of highly educated speakers. This effect may be due to the fact that in Flanders 'standard Dutch' is more formal than in the Netherlands, and therefore more 'informational' which in turn leads highly educated speakers to target more infrequent and new words. We did not find any educational effects for the proportion of MCWs, nouns, adjectives and verbs speakers used. The higher productivity figures of highly educated speakers in public speech match the higher productivity figures in the quality newspapers, aiming at a highly educated readership.

6.2.5 Age

The speaker's age also turned out to be a substantial predictor of lexical variation. The speaker's lexical knowledge and creativity appear to increase during the lifetime, under the condition that the speaker was exposed to a lexically

rich environment, as is more common for highly educated speakers. This interpretation is based on our analyses of derivational and lexical productivity, where we observed that older, highly educated speakers used their lexicon more productively. Middle aged and old speakers use their capacity to use their language resources more productively mostly in situations that evoke the use of more 'informational' language. In private spontaneous dialogues derivational productivity was comparably low for all age groups. Lexical productivity, however, was somewhat higher for older speakers. However, it was much lower than in public dialogues, in which productivity for middle-aged and old speakers was much higher. Given these results, it was not surprising not to find effects for age in Chapter 5 where we only investigated spontaneous private dialogues. However, we did find a higher proportion of nouns for older speakers in this chapter, which also points at the expansion of lexical knowledge over lifetime. We also found that young speakers used adjectives most frequently. However, this effect is word-bound, and solely the effect of a few adjectives typically used by young speakers as popular interjections and discourse markers. This result confirms the outcomes of Rayson et al. (1997) who observed that young speakers more intensively use specific adjectives. In the Netherlands old speakers use the language most productively, while in Flanders, middle age speakers are the most productive age group. This may be an effect of the specific standardization process of the Dutch language in Flanders where the more prestigious Netherlandic Dutch variety was adopted in 1898, and only became the single official language in Flanders in 1930 (for more details see Grondelaers and Van Hout, 2011).

6.3 Discussion and future research

This dissertation presents an overview of the register and sociolinguistic variation observed in lexical patterns across spoken and written Dutch. We specifically targeted lexical productivity and derivational morphology. We always started by exploring global patterns and, if possible, we continued by taking a closer look at the specific or autonomous contributions of specific affixes and words. Many word and affixes turned out to have their own variation patterns, implying that often no overarching, global patterns could be established. We found no general consistent pattern describing the variation of all words ending in the suffix *-lijk*. This suggests that specific lexical characteristics cannot be described with rule-based linguistic principles only. The relations between these words and affixes must be explained primarily on word-specific patterns and effects. This conclusion opens the door for exemplar-based approaches. Exemplar theorists claim that memory-stored pieces of information such as words and idioms are the building blocks for language structure, created by analogical generalizations over stored chunks (Bresnan and Hay, 2008: 256). This means that further research on lexical patterns and their sources of variation really could contribute to the development of exemplar-based theories on linguistic

variation.

There are many patterns of variation in both lexical characteristics and individual words that may remain unnoticed in explorations of register and sociolinguistic variation, given the corpus sizes and the many different words in these corpora. In Chapter 5 we demonstrated the power of random sampling corpora, in fact a method comparable to bootstrapping in statistics. For language and speech corpora, it has the advantage of removing the dependencies between sequences of words. The samples contain elements that were drawn independently. The counts and measurements we used to analyze different linguistic categories were all normally distributed, which made the statistical analysis and results more easy to perform and interpret. Schoonewelle (2011) carried out a further test of this approach by using these random samples to explore sociolinguistic variation in the use of nine lexical categories in spontaneous Dutch speech. Her main aim was to reveal gender differences, which she found in the use of articles, negations, pronouns, and words expressing negative emotions. She also included the social factors age and education level in her analyses. The speaker's age appeared to be a highly influential factor in the use of many of the lexical categories investigated whereas education level only occasionally emerged as an influential predictor. The results of Schoonewelle (2011) corroborate the potential power of repeated random sampling for lexical research.

There is at the same time a disadvantage to using samples. Data remain unused, which makes exploration of lexical characteristics with a low frequency, such as derivational productivity, difficult, if not impossible. It turned out that is was feasible to work with the complete subcorpora, even though they varied enormously in size. Although we worked with sparse data in these analyses, fitting a generalized linear model with a binomial link function, predicting successes and failures (for instance the word being a derivational hapax or not), gave reliable results. Principal Component Analysis proved to be a useful first step to obtain a global overview of the data, regardless of working with complete subcorpora or data samples. When in the future even larger speech corpora become available marked by a more equal distribution of speech over different social groups, we can construct more and larger data samples, making it possible to investigate many, less frequently occurring, lexical characteristics on the basis of randomly selected data samples. This applies to the exploration of derivational productivity, but also for the exploration of family words (for instance 'mom', 'brother'), swear words, and sensations (for instance, 'feel', 'listen', 'view'), for which in the analyses of Schoonewelle (2011) not enough data was available.

The rise of the new media, such as chat and twitter, makes it possible to obtain and analyze huge corpora containing informally written texts. Chat language has many characteristics comparable to private spontaneous speech: It is mostly used in informal, private settings, and the writer has limited time to reconsider his or her lexical choices. The study of large corpora of written informal language offers new ways of coming to grips with how language works,

how language is connected to social and emotional meaning, and how different alternative expressions interact and compete in language production. One of the characteristics of more informal writing is to adapt the written format to the actual speech form. A quick search on *tuuk* ('natuurlijk', meaning 'of course') on the web returned, among others, a Tweet (Twitter message) containing two highly reduced word forms of words ending in the suffix *-lijk*, namely *tuuk* and *eik* (*eigenlijk*, meaning 'actually'):

> *Sanne Wallis Devries zegt: @caricevhouten Eerst: 'wat dan?'*
> *en toen: 'wat dahan?!'..maar nu is 't eik al nie meer leuk, tuuk.*
> (Sanne Wallis Devries says: @caricevhouten First: 'what then?'
> and then: 'what thehen?!'..but now it is actually not funny anymore, of course.)
> `http://www.bnerslive.nl/160335547536113664)` (May 2012)

Our attempt to combine the fields of sociolinguistics and corpus linguistics in studying lexical variation in two large corpora proved to be successful. Both language *use*, defined by register, and the language *user*, defined by the sociolinguistic variables, emerged as important sources of lexical variation. We even disclosed important interactions between register and the sociolinguistic variables, namely that effects for the sociolinguistic variables 'education level' and 'age' were only present in public speech and not in private speech. The inclusion of registers widens the sociolinguistic scope enormously, which is necessary to expand the field of language variation and change to other domains of language use and to overcome the narrow focus on spontaneous speech data only. The definition of the sociolinguistic style continuum on the one-dimensional basis of attention paid to speech only is too limited (cf. Eckert and Rickford, 2011) to develop further insights into patterns of lexical variation. The same can be said about the variable rule analysis. We need more techniques in studying variation in language use and language users, as we have shown, to get a deeper understanding of variation in language.

References

Biber, D. and S. Conrad, 2009. *Register, Genre, and Style*. Cambridge University Press, Cambridge

Biber, D., S. Conrad and R. Reppen, 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, Cambridge

Bresnan, J. and J. Hay, 2008. An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua*, 118: 245–259

Burrows, J. F., 1992. Computers and the study of literature. In C. S. Butler, ed., *Computers and Written Texts*. Blackwell, Oxford, 167–204

Burrows, J. F., 1993. Tiptoeing into the infinite: Testing for evidence of national differences in the language of English narrative. In S. Hockey and N. Ide, eds., *Research in Humanities Computing '92*. Oxford University Press, London

Coates, J., ed., 1998. *Language and gender: A Reader*. Blackwell, Oxford

Eckert, P. and J. R. Rickford, eds., 2011. *Style and Sociolinguistic Variation*. Cambridge University Press, Cambridge

Grondelaers, S., K. Deygers, H. van Aken, V. van den Heede and D. Speelman, 2000. Het CONDIV-corpus geschreven Nederlands. *Nederlandse Taalkunde*, 5: 356–363

Grondelaers, S. and R. van Hout, 2011. The standard language situation in the Low Countries: Top-down and bottom-up variations on a diaglossic theme. *Journal of Germanic Linguistics*, 23 (3): 199–243

Härnqvist, K., U. Christianson, D. Ridings and J.-G. Tingsell, 2003. Vocabulary in interviews as related to respondent characteristics. *Computers and the Humanities*, 37: 179–204

Keune, K., S. van Gijsel, R. van Hout and R. H. Baayen, submitted. Sociolinguistic patterns in dutch: Measuring lexical characteristics of spontaneous speech. Speech communication

Keune, K., R. van Hout and R. H. Baayen, 2006. Socio-geographic variation in morphological productivity in spoken dutch: A comparison of statistical techniques. In J.-M. Viprey, ed., *Actes des 8es journées internationales d'analyse statistique des données textuelles*, volume 2. Presses Universitaires de Franche-Comté, 571–580

Newman, M. L., C. J. Groom, L. D. Handelman and J. W. Pennebaker, 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45: 211–236

Oostdijk, N., W. Goedertier, F. van Eynde, L. Boves, J. P. Martens, M. Moortgat and R. H. Baayen, 2002. Experiences from the Spoken Dutch Corpus Project. In M. Gonz ez Rodriguez and C. Paz Su ez Araujo, eds., *Proceedings of the third International Conference on Language Resources and Evaluation*. 340–347

Plag, I., C. Dalton-Puffer and R. H. Baayen, 1999. Morphological productivity across speech and writing. *English Language and Linguistics*, 3 (2): 209–228

Pluymaekers, M., M. Ernestus and R. H. Baayen, 2005. Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, 118(4): 2561–2569

Rayson, P., G. Leech and M. Hodges, 1997. Social differentiation in the use of English vocabularly: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2 (1): 133–152

Schoonewelle, A., 2011. Genderverschillen in het gesproken Nederlands: Een corpusstudie. Bachelors thesis, Radboud University Nijmegen

Van Gijsel, S., 2007. Sociovariation in Lexical Richness. A Quantitative Corpus Linguistic Analysis. Ph.D. thesis, Katholieke Universiteit Leuven

Samenvatting

Teksten bevatten een grote verscheidenheid aan lexicale elementen. Er is bijvoorbeeld variatie in het gebruik van zelfstandige naamwoorden, adjectieven en werkwoorden, in het aantal gebruikte en de soort van samengestelde woorden, in de mate van gebruik van hoogfrequente woorden en functiewoorden en in het aantal unieke woorden in een tekst. Dit proefschrift gaat over de invloed van register en sociolinguïstische factoren op de manier waarop lexicale karakteristieken variëren. Het begrip register (Reid, 1956; Halliday, 1964; Biber, 1988, 1995; Biber en Conrad, 2009) verwijst naar de verschillende manieren waarop taal gebruikt kan worden. Register omvat de functionele variatie die in principe beschikbaar is voor elke taalgebruiker in tegenstelling tot sociolinguïstische elementen die juist verwijzen naar de karakteristieken van de specifieke taalgebruiker.

De lexicale elementen die in een krantenbericht gebruikt worden zullen duidelijk anders zijn dan de lexicale elementen in een informeel e-mailbericht. Een krantenbericht zal meer lexicale elementen bevatten die informatief van aard zijn, terwijl een informeel e-mailbericht door de bank genomen meer lexicale elementen zal bevatten die persoonlijke betrokkenheid uitdrukken. Een krantenbericht en een e-mailbericht zijn dus twee verschillende registers in geschreven taal. Ook binnen krantenberichten kan onderscheid gemaakt worden tussen verschillende registers. Kwaliteitskranten als het NRC Handelsblad richten zich op een ander lezerspubliek dan tabloids als de Telegraaf. Het valt te verwachten dat de taal die ze gebruiken, inclusief de keuze van lexicale elementen, afgestemd zal zijn op het lezerspubliek. Registerverschillen zijn er vanzelfsprekend niet alleen in geschreven, maar ook in gesproken taal. In een voorbereide presentatie zullen andere lexicale elementen de voorkeur krijgen dan in een spontaan telefoongesprek (dialoog). De volgende sociolinguïstische factoren zij meegenomen in het onderzoek: het land van herkomst (Nederland versus Vlaanderen), het geslacht, het opleidingsniveau (hoog versus niet-hoog) en de leeftijd van de taalgebruiker (< 40, $40 - 60$, > 60 jaar).

Waar corpuslinguïsten zich voornamelijk richten op de effecten van register op globale lexicale tekstkarakteristieken, ligt de nadruk bij variatieonder-

zoek in de sociolinguïstiek juist op de invloed van sociale factoren op specifieke linguïstische kenmerken ofwel variabelen. Deze zijn meestal fonologisch of morfo-syntactisch van aard. In dit onderzoek hebben we getracht de corpuslinguïstiek en de sociolinguïstiek te combineren door globale lexicale variatiepatronen te bestuderen op basis van twee grote corpora.

Om globale variatiepatronen te onderzoeken in het geschreven Nederlands hebben we gebruik gemaakt van het CONDIV corpus (Grondelaers et al., 2000). Dit corpus bevat ongeveer 17,6 miljoen woorden afkomstig uit drie Nederlandse en vier Vlaamse kranten. Zowel voor Nederland als voor Vlaanderen bevat het krantenartikelen uit een kwaliteitskrant, een tabloid met een nationale verspreiding en een regionale krant (voor Vlaanderen twee), wat het mogelijk maakt om onderscheid te maken tussen verschillende registers in Nederland en Vlaanderen.

Voor het gesproken Nederlands hebben we taal uit het Corpus Gesproken Nederlands (CGN) (Oostdijk et al., 2002) onderzocht. Dit corpus bevat ongeveer 8,9 miljoen woorden uit spraakfragmenten van Nederlandse en Vlaamse volwassenen, met spraak uit verschillende registers. In ons onderzoek hebben we ons beperkt tot private spraak (spontane conversaties en telefoongesprekken: 4,7 miljoen woorden) en publieke spraak (3,4 miljoen woorden). De publieke spraak kan opgesplitst worden in twee categorieën: dialogen (bijv. debatten, vergaderingen en interviews) en monologen (bijv. nieuws, commentaren en lessen). Doordat het CGN sprekerinformatie bevat, was het mogelijk de factoren land, geslacht, opleidingsniveau en leeftijd te onderscheiden.

In Hoofdstuk 2 hebben we de invloed van register en land op de frequentie van het gebruik van woorden eindigend op het suffix *-lijk* onderzocht. Dit suffix is nauwelijks meer productief (geen nieuwe vormen) en kan als gelexicaliseerd beschouwd worden. Het woord *natuurlijk* betekent tegenwoordig meestal 'vanzelfsprekend' en heeft zijn oorspronkelijke betekenis 'zoals in de natuur' verloren. Door verlies aan inhoudelijke informatie en de lexicalisatie van de woordvorm zijn verscheidene van deze woorden onderhevig aan sterke akoestische reductie in hun uitspraak: *natuurlijk* wordt vaak gereduceerd tot *ntuuk* en zelfs tot *tuuk*. Om onze hypothese te testen dat de variatie in het gebruik van een specifieke lexicale categorie, nl. de woorden eindigend op *lijk*, ook aanwezig is in andere delen van het lexicon, creëerden we een referentiepunt. We onderzochten de invloed van register en land op het gebruik van de 'most common words' (de meest gebruikelijke woorden, ofwel de meest frequente woorden, MCWs), die de syntactische gewoontes van taalgebruikers aanboren (Burrows, 1992, 1993). In het geschreven Nederlands vonden we dezelfde patronen voor register en land in beide lexicale categorieën. In het gesproken Nederlands vonden we evenwel sociolinguïstische verschillen: hoogopgeleide sprekers gebruikten woorden eindigend op het suffix *-lijk* frequenter dan niet-hoogopgeleide sprekers en vrouwen gebruiken woorden met *-lijk* frequenter dan mannen. Dit resultaat suggereert dat de beide lexicale categorieën uit onafhankelijke bronnen putten. Verder vonden we dat het aantal voorkomens van de afzonderlijke woorden eindigend

op *-lijk* en ook van de MCWs varieerde afhankelijk van elk van de predictoren in zowel geschreven als gesproken Nederlands. In het gesproken Nederlands waren de woorden *tamelijk* en *ongelooflijk* typerend voor mannen terwijl de woorden *vriendelijk* en *lelijk* meer typerend voor vrouwen waren. Deze woordspecifieke verschillen indiceren dat een lexicale categorie niet uit een coherente, onderling samenhangende set woorden bestaat, maar dat de betrokken woorden hun eigen lexicale distributie hebben. Vervolgens hebben we de mate van akoestische reductie van woorden eindigend op *-lijk* onderzocht. We vonden dat mannen sterker reduceerden dan vrouwen en dat in Vlaanderen hoogopgeleide sprekers minder gereduceerde vormen gebruikten dan niet-hoogopgeleide sprekers. Tussen Nederland en Vlaanderen vonden we verder geen globale verschillen, maar wel woordspecifieke reductiepatronen. De mate van reductie werd mede bepaald door twee linguïstische factoren, namelijk de positie van het woord in de zin (finaal/niet-finaal) en de mate van voorspelbaarheid van het woord op grond van de context. Woorden in zinsfinale positie vertoonden minder reductie en woorden met een hoge voorspelbaarheid op basis van het voorafgaande woord vertoonden juist meer reductie. Dit suggereert dat er fundamentele linguïstische factoren zijn die onafhankelijk van registers, landen en sociale factoren opereren.

In Hoofdstuk 3 hebben we regionale (land) en sociale (geslacht, opleidingsniveau en leeftijd) variatiepatronen onderzocht voor potentiële productiviteit ofwel de verwachte mate van toename van derivationele vormen in het spontaan gesproken Nederlands en hebben we onderzocht met welke statistische techniek onze data het best geanalyseerd kunnen worden. Hiervoor hebben we voor 72 affixen het aantal hapax legomena (woorden die slechts eenmaal in het corpus voorkomen) geselecteerd uit de subcorpora van het Corpus Gesproken Nederlands die uit private en publieke spontane spraak bestaan. Vervolgens hebben we de distributie van deze hapax legomena over de 24 subcorpora, die we gedefinieerd hebben door onderscheid te maken naar land, geslacht, opleidingsniveau en leeftijd ($2 \times 2 \times 2 \times 3$), geanalyseerd. Doordat er voor veel affixen geen voorkomens waren in een of meerdere subcorpora hadden we te maken met veel cellen met nulwaarden hetgeen het extra lastig maakt om een passend model te vinden voor de analyse van de data. We vergeleken drie statistische technieken: het gewone kleinste-kwadratenmodel (OLS) met de getransformeerde proportie hapax legomena in het subcorpus als de afhankelijke variabele, een lineair mixed effects model met affix als random effect en wederom de getransformeerde proporties hapax legomena in het subcorpus als afhankelijke variabele en een generalized linear model met een binomiale linkfunctie (logit) waarin de hapax legomena als successen en de overige woorden als niet-successen werden beschouwd. De laatstgenoemde techniek gaf zeer goede resultaten. Het omgaan met de cellen met nulwaarden en het omgaan met hapax legomena, die een Zipfiaanse distributie hebben, verliep met deze technniek het beste. We vonden met alledrie de technieken resultaten die in dezelfde richting wezen: hoogopgeleide oudere mannen vertoonden de hoogste productiviteit. Ook toonden de modellen aan

dat er geen globaal regionaal en sociaal patroon is. Wel waren er wederom veel affix-specifieke verschillen.

In Hoofdstuk 4 hebben we de effecten van register en sociolinguïstische factoren op derivationele en lexicale productiviteit in zowel het geschreven (CONDIV krantencorpus) als het gesproken Nederlands (CGN) onderzocht. Net als in Hoofdstuk 3 hebben we de distributie van het totale aantal (derivationele) hapax legomena over de subcorpora geanalyseerd. Onze hypothese dat derivationele productiviteit lexicale productiviteit weerspiegelt, bleek juist te zijn. Lexicale productiviteit bleek hoger te zijn in geschreven dan in gesproken Nederlands Dit is in lijn met eerder onderzoek (cf. Biber en Conrad, 2009; Plag et al., 1999). We vonden ook een hogere productiviteit in de formelere registers binnen het geschreven en gesproken Nederlands. Derivationele productiviteit bleek echter niet hoger in geschreven dan in gesproken Nederlands. Tussen Nederland en Vlaanderen vonden we wederom geen globaal effect. Er waren wel affixspecifieke verschillen. Alle sociale factoren bleken van invloed op de productiviteit: hoogopgeleide oudere mannen waren het productiefst in het gebruik van zowel derivationele als lexicale items. Het effect voor leeftijd kwam niet naar voren bij niet-hoogopgeleide sprekers en de private telefoondialogen.

In Hoofdstuk 5 hebben we sociolinguïstische patronen in globale lexicale variatie in spontane spraak uit het Corpus Gesproken Nederlands onderzocht. We hebben met drie maten gewerkt die globale lexicale variatie meten: lexicale diversiteit (gemeten door het aantal types en hapax legomena te tellen), lexicale densiteit (gemeten door het aantal zelfstandige naamwoorden, adjectieven en werkwoorden te tellen) en lexicale communaliteit (gemeten door de voorkomfrequentie van de most common words te bepalen). Om effecten van tekstlengte in de maten die lexicale diversiteit meten te vermijden, hebben we ervoor gekozen om met steekproeven te werken die elk evenveel woorden bevatten. We vergeleken de effecten van gerandomiseerde met niet-gerandomiseerde steekproeven (voor telkens 10 steekproeven van 1350 woorden) en het bleek dat de gerandomiseerde steekproeven (de toevalssteekproeven) veel meer statistisch vermogen (power) hadden en superieur waren in het traceren van sociolinguïstische patronen. Principale componenten analyse bleek een geschikte techniek om een globaal beeld van de data te krijgen. Om meer informatie over de significantie, effectgrootte en de mogelijke interacties in de data te krijgen hebben we de data met general linear models geanalyseerd. Geslacht bleek de variabele te zijn die de meeste variatie verklaart. Dit effect was globaal, dus niet afhankelijk van een aantal specifieke woorden. Mannen gebruikten meer hapax legomena, meer types en meer zelfstandige naamwoorden, vrouwen meer werkwoorden en MCWs. Deze resultaten zijn in overeenstemming met eerder onderzoek (Newman et al., 2008; Härnqvist et al., 2003; Rayson et al., 1997; Coates, 1998). Nederlandse sprekers gebruikten meer MCWs en adjectieven, terwijl Vlaamse sprekers meer werkwoorden gebruikten. Alle effecten voor land bleken woordgebonden en waarschijnlijk voort te komen uit divergente, lexicale keuzes om specifieke concepten uit te drukken. Oudere sprekers gebruikten de meeste zelf-

standige naamwoorden (globaal effect), wat mogelijk een effect is van de verdere ontwikkeling van de lexicale kennis gedurende het leven. Verder vonden we nog een duidelijk woordgebonden effect voor het frequente gebruik van adjectieven onder jongeren. Dit werd veroorzaakt door het intensieve gebruik van specifieke affixen, populaire interjecties en discourse markers.

Wat voor registereffecten hebben we gevonden en wat is de invloed van de onderzochte sociolinguïstische variabelen als we alle resultaten bij elkaar leggen?

Register

Het onderscheid tussen het geschreven en gesproken Nederlands bleek een belangrijke voorspeller van variatiepatronen in het lexicon te zijn. Zowel in geschreven als in gesproken Nederlands vonden we de hoogste derivationele en lexicale productiviteit in de formeelste registers. We vonden ook een hogere globale lexicale productiviteit in het geschreven dan in het gesproken Nederlands. Deze uitkomsten zijn overeenkomstig het 'informational versus involved' criterium dat Biber en Conrad (2009) identificeerden om geschreven en gesproken taal te contrasteren. Voor derivationele productiviteit vonden we echter geen hogere productiviteit in het geschreven Nederlands. De productiviteit van de individuele affix varieerde sterk binnen het geschreven en gesproken Nederlands: sommige affixen bleken typisch voor het geschreven Nederlands en andere voor het gesproken Nederlands. Een mogelijke verklaring voor de hoge productiviteit van sommige affixen in spontane spraak is dat sprekers actiever gebruik maken van de productieve eigenschappen van affixen om nieuwe woorden te creëren. Ze hebben niet veel tijd om lexicale keuzes te overwegen en overdenken en zijn daardoor geneigd productieve affixen te gebruiken om lexicale beslissingen te faciliteren.

Nederland versus Vlaanderen

Het belangrijkste resultaat dat we hebben gevonden met betrekking tot de verschillen tussen Nederland en Vlaanderen is dat de variatiepatronen voornamelijk woord- of affixgebonden zijn. Verschillen in de frequentie van gebruik van woorden eindigend op *-lijk* en de reductie van deze woorden, verschillen in derivationele en lexicale productiviteit en ook verschillen in frequentie van het aantal adjectieven, werkwoorden en MCWs bleken allen woordgebonden te zijn. Deze verschillen komen waarschijnlijk voort uit divergente lexicale keuzes om specifieke concepten uit te drukken. Dit komt overeen met eerder onderzoek waarin wordt gesteld dat, met uitzondering van een aantal specifieke hoogfrequente karakteristieken van 'Tussentaal' (zie ook Hoofdstuk 5), het Nederlands zoals het gesproken wordt in Vlaanderen niet fundamenteel verschilt van het Nederlands zoals het in Nederland gesproken wordt (cf. Grondelaers en Van Hout, 2011).

Geslacht

Geslacht bleek, samen met register, de sterkste voorspeller van globale lexicale variatiepatronen. De variatiepatronen waren in tegenstelling tot die voor land voornamelijk globaal: uitsluiting van een aantal woorden die specifiek voor mannen of vrouwen leken te zijn, veranderde de resultaten niet. Dit sluit aan bij het systematische effect voor geslacht zoals dit in sociolinguïstisch onderzoek gevonden wordt (Coates, 1998). Een hoge derivationele en lexicale productiviteit, een hoge Type-Token Ratio en het gebruik van veel zelfstandige naamwoorden, allen kenmerkend voor een meer 'informatieve' gesprekssstijl, is karakteristiek voor spraak van mannen (Biber en Conrad, 2009; Rayson et al., 1997). Het gebruik van een groot aantal werkwoorden en MCWs, kenmerkend voor een meer 'betrokken' gesprekssstijl, is karakteristiek voor spraak van vrouwen. Deze uitkomsten zijn in overeenstemming met uitkomsten uit eerder onderzoek van Newman et al. (2008) en Härnqvist et al. (2003). Inspectie van individuele woorden die naar voren kwamen als typisch voor mannen of typisch voor vrouwen bevestigden onze conclusie dat mannen een meer 'informatieve' gesprekssstijl hebben en vrouwen zich op een meer 'betrokken' manier uitdrukken. Het verschil tussen mannen en vrouwen lijkt voornamelijk een sociaal-cultureel effect te zijn. Vrouwen hanteren een meer 'betrokken' gesprekssstijl dan mannen omdat 'betrokken' spraak in onze samenleving als meer gepast wordt beschouwd voor vrouwen. Het is wellicht interessant om te speculeren dat een meer 'betrokken' gesprekssstijl voor vrouwen biologisch bepaald is, aangezien vrouwen onafhankelijk van de tijd en cultuur waarin ze leven, meestal een intensievere taak hebben in de opvoeding van kinderen, maar voor een dergelijke claim is aanvullend interdisciplinair onderzoek nodig.

Opleidingsniveau

Het opleidingsniveau van een spreker bleek zowel voorspellend voor de mate van productiviteit in de spraak als voor de frequentie waarmee de spreker woorden eindigend op *-lijk* gebruikt. Kennelijk zijn er bepaalde vaardigheden nodig om nieuwe woorden te produceren of om dieper uit lexicale bronnen te kunnen putten, die ruimer beschikbaar zijn voor hoogopgeleide dan voor niet-hoogopgeleide sprekers. Deze resultaten komen overeen met resultaten uit onderzoek van Härnqvist et al. (2003). In de private spontane dialogen kwam opleidingsniveau in Nederland niet naar voren als een voorspeller van lexicale variatie. Dit is niet verrassend, aangezien sprekers uitgedaagd moeten worden om productievere, lexicaal rijkere taal te gebruiken. Aangezien private spontane spraak over het algemeen meer 'betrokken' en minder 'informatief' is, is het waarschijnlijk dat hoogopgeleide sprekers niet hun maximale capaciteit benutten om infrequentere en nieuwe woorden te gebruiken. In Vlaanderen vonden we wel een effect voor opleidingsniveau in private spontane spraak. Dit kan een gevolg zijn van het feit dat Standaardnederlands in Vlaanderen formeler is dan in Nederland en daardoor 'informatiever' dan in Nederland.

Leeftijd

Ook de leeftijd van de spreker kwam naar voren als een belangrijke voorspeller van lexicale variatie. Het blijkt dat de lexicale kennis en creativiteit van de spreker toeneemt gedurende zijn of haar leven, onder de voorwaarde dat de spreker blootgesteld is aan een lexicaal rijke omgeving, wat gebruikelijker is voor hoogopgeleide sprekers. Deze interpretatie is gebaseerd op onze analyses van derivationele en lexicale productiviteit. Ook het hogere aandeel zelfstandige naamwoorden in spraak van ouderen wijst op de toename van lexicale kennis gedurende het leven. Het hoge aandeel adjectieven in het spraakgebruik van jongeren bleek veroorzaakt door het intensieve gebruik van een aantal specifieke woorden die gebruikt worden als populaire interjecties en discourse markers. In Nederland kwam de oudste groep sprekers naar voren als het meest productief, terwijl in Vlaanderen de sprekers van 40 tot 60 jaar het productiefst waren. Dit verschil is mogelijk te verklaren door het specifieke standaardisatieproces van het Nederlands (voor meer details zie Grondelaers en Van Hout, 2011).

Onze poging om de sociolinguïstiek en corpuslinguïstiek te combineren in het bestuderen van twee grote corpora bleek succesvol te zijn. Zowel het *taalgebruik* (register) als de kenmerken van de *taalgebruiker* (sociolinguïstische variabelen) bleken belangrijke voorspellers van lexicale variatie te zijn. Het toevoegen van registers aan sociolinguïstisch onderzoek verbreedt het sociolinguïstische blikveld en is nodig om onderzoek in het veld uit te breiden naar nieuwe dimensies van taalvariatie en taalverandering.

Discussie en toekomstig onderzoek

Dit onderzoek bevat een overzicht van lexicale variatiepatronen. Veel woorden en affixen blijken hun eigen variatiepatronen te hebben. Dit suggereert dat specifieke lexicale eigenschappen niet met regelgebaseerde linguïstische principes beschreven kunnen worden, maar dat de relaties tussen affixen en woorden voornamelijk door woordspecifieke patronen en effecten beschreven moeten worden. Deze conclusie opent de deur voor een zogeheten exemplar-based aanpak. In die benadering wordt geclaimd dat stukjes informatie, zoals woorden en idiomen, rechtstreeks in het geheugen opgeslagen worden en dat ze vervolgens als bouwstenen gebruikt worden voor de constructie van taalstructuur, op grond van analogie (Bresnan en Hay, 2008: 256). Dit betekent dat verder onderzoek naar lexicale patronen en de bijbehorende bronnen van variatie wezenlijk zou kunnen bijdragen aan de ontwikkeling van exemplar-based theorieën over lexicale variatie. Er zijn veel variatiepatronen, zowel in lexicale categorieën als in afzonderlijke woorden, die onzichtbaar blijven in dit onderzoek door de beperkte omvang van subcorpora en de vele verschillende woorden in de corpora. In Hoofdstuk 5 hebben we de kracht van het werken met toevalssteekproeven laten zien, een methode die in feite vergelijkbaar is met het gebruik van bootstrapping in de statistiek. Er is echter ook een nadeel aan het gebruik van toevalssteekproeven: beschikbare data blijven ongebruikt, wat onderzoek

naar lexicale karakteristieken die laagfrequent zijn moeilijk of zelfs onmogelijk maakt. Uit Hoofdstuk 3 en Hoofdstuk 4 blijkt dat het mogelijk is om voor het onderzoek naar derivaties, die vaak laagfrequent zijn, met alle beschikbare data te werken. Het toepassen van een generalized linear model met een binomiale linkfunctie, waarin successen en niet-successen voorspeld werden (bijvoorbeeld of het woord een derivationele hapax is of niet), gaf betrouwbare resultaten. Principale componenten analyse blijkt een betrouwbare eerste stap om een overzicht over de data te krijgen. De opkomst van nieuwe media als chat en twitter maakt het mogelijk om enorme corpora bestaande uit informeel geschreven Nederlands te creëren en te onderzoeken. Chattaal heeft veel eigenschappen die vergelijkbaar zijn met private spontane spraak: het wordt meestal gebruikt in privésituaties en de spreker heeft zeer beperkte tijd om zijn of haar lexicale keuzes te overwegen. Het bestuderen van grote corpora, bestaande uit informele taal, geeft de mogelijkheid om beter te begrijpen hoe taal in elkaar zit, hoe de taal verbonden is met sociale en emotionele betekenis en hoe verschillende alternatieve expressies interacteren in taalproductie.

Referenties

Biber, D., 1988. *Variation Across Speech and Writing.* Cambridge University Press, Cambridge

Biber, D., 1995. *Dimensions of Register Variation.* Cambridge University Press, Cambridge

Biber, D. en S. Conrad, 2009. *Register, Genre, and Style.* Cambridge University Press, Cambridge

Bresnan, J. en J. Hay, 2008. An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua*, 118: 245–259

Burrows, J. F., 1992. Computers and the study of literature. In C. S. Butler, ed., *Computers and Written Texts.* Blackwell, Oxford, 167–204

Burrows, J. F., 1993. Tiptoeing into the infinite: Testing for evidence of national differences in the language of English narrative. In S. Hockey en N. Ide, eds., *Research in Humanities Computing '92.* Oxford University Press, London

Coates, J., ed., 1998. *Language and gender: A Reader.* Blackwell, Oxford

Grondelaers, S., K. Deygers, H. van Aken, V. van den Heede en D. Speelman, 2000. Het CONDIV-corpus geschreven Nederlands. *Nederlandse Taalkunde*, 5: 356–363

Grondelaers, S. en R. van Hout, 2011. The standard language situation in the Low Countries: Top-down and bottom-up variations on a diaglossic theme. *Journal of Germanic Linguistics*, 23 (3): 199–243

Halliday, M. A. K., 1964. Comparison and translation. In M. Halliday, M. McIntosh en P. Strevens, eds., *The linguistic sciences and language teaching*. Longman, London

Härnqvist, K., U. Christianson, D. Ridings en J.-G. Tingsell, 2003. Vocabulary in interviews as related to respondent characteristics. *Computers and the Humanities*, 37: 179–204

Newman, M. L., C. J. Groom, L. D. Handelman en J. W. Pennebaker, 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45: 211–236

Oostdijk, N., W. Goedertier, F. van Eynde, L. Boves, J. P. Martens, M. Moortgat en R. H. Baayen, 2002. Experiences from the Spoken Dutch Corpus Project. In M. Gonz ez Rodriguez en C. Paz Su ez Araujo, eds., *Proceedings of the third International Conference on Language Resources and Evaluation*. 340–347

Plag, I., C. Dalton-Puffer en R. H. Baayen, 1999. Morphological productivity across speech and writing. *English Language and Linguistics*, 3 (2): 209–228

Rayson, P., G. Leech en M. Hodges, 1997. Social differentiation in the use of English vocabularly: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2 (1): 133–152

Reid, T. B., 1956. *Linguistics, Structuralism, Philology*, *Archivum Linguisticum*, volume 8. Jackson, Son & Company, Glasgow

Acknowledgements

Finally, after reading the acknowledgements of many of my (former) fellow students, my time has come to express my gratitude to the many people who supported me during my PhD project.

First of all, I am indebted to my supervisors Roeland van Hout and Harald Baayen for their invaluable guidance. Thank you Harald, for your incredible enthusiasm, inspiration and support during the first years of my PhD project. Your passion for data analysis has been contagious. Roeland, without your help in the last, more difficult, years of my PhD project, I would not have reached the point of writing these acknowledgements. Thank you very much for taking over the supervision of my PhD project after Harald moved to Alberta. You never gave up on me, not after my pregnancies during which I was not able to work, and not in the period in which I was sometimes less inspired as an effect of my many sleepless nights. Thank you for sharing your linguistic and statistical knowledge with me at our weekly meetings and for helping me during the writing process. The third person that has been very important to me and who has had a constant presence throughout my PhD project is Mirjam Ernestus. Mirjam, thank you very much for your confidence and interest in my work. I highly value the collaboration on my first article, your useful and critical comments on my presentations and, of course, all the conversations, related and unrelated to my research.

I would also like to thank the members of my manuscript committee, Margot van Mulken, John Nerbonne and Sally Tagliamonte, who made time to read my manuscript before the summer holidays.

Furthermore, I owe many thanks to all my colleagues at the CLSM (formerly known as the IWTS), Anneke, Anneli, Arina, Berit, Debby, Esther, Femke, Francisco, Hanke, Harald, Hilde, Inge, Iris, Jeanne, Kors, Lanneke, Laura, Luuk, Marco, Marieke, Marjolein, Mark, Mirjam, Mybeth, Patricia, Rob, Roel, Stefan, Victor, Wieke and many others, who made the CLSM a very inspiring and pleasant working environment for me. Thank you all for the stimulating discussions and valuable lunch breaks. I really enjoyed the numerous social activities we organized in my first years as a PhD student: from having drinks and

stamppot in the colloquium room to late nights out in Nijmegen and the tour through Maastricht led by our own official tour guide Mark. Rob Schreuder, I really appreciate your kind support, interest and trust in me and my work. Laura, Iris, Marco and Francisco, thank you for being such nice roommates. Special thanks go to Wieke and Mark. Wieke, the combination of sharing so many things being roommates for years and our many common interests led to a close friendship. I really miss our never-ending conversations now you live near San Francisco. I look forward to spending more time together when you get back to the Netherlands. Mark, thanks for being such a great roommate and friend. Thank you for being my 'human agenda' and always reminding Wieke and me of the advantages of having our stuff better organized. You were always right. With you as my *paranimf*, I am sure nothing can go wrong.

I would like to thank *de Leuvenaren*, who also took part in this NWO project, Dirk Speelman, Dirk Geeraerts, Stef Grondelaers, Koen Plevoets, and Sofie van Gijsel for their stimulating discussion. Sofie, it was always a pleasure working with you. I am glad we decided to work together on the third article in this dissertation. Thank you for your great company during the conference in Besançon.

I cannot look back on my years as a PhD student without thinking of the many colleagues who gave me a great social life in the first years of my PhD project. Else, Femke, Iris, Josje, Suzanne, and Wieke, thank you for our nice get-togethers. PhD students from the *Erasmusgebouw*, among others, Nienke, Peter, Rik, and Folkert, thank you for all the *borrels* in the Cultuurcafé and for the dinners afterwards. Annika, what a coincidence we had the same supervisor and were trying hard to finish our PhD while living in Duiven, raising our children and having a new job at the same time. It is great to share these experiences with you.

I owe many thanks to my new colleagues at Cito, who gave me the opportunity to take some time off in the last year to finish this dissertation. Thank you for your support and understanding.

To my friends and family I would like to say thank you for your patience and understanding when I was too busy to visit you often in the last couple of years. I promise I will do better next year. Reanne, thank you for your friendship and the many nice mini holidays. I am happy you agreed to be my *paranimf*.

As promised to my oldest son Stan, I will now continue in Dutch: pap, mam, Anne en Merlijn, heel fijn dat jullie er altijd voor me zijn. Pap, mam, zonder jullie steun en goede zorg voor de kinderen in de afgelopen jaren was het nooit gelukt mijn proefschrift af te ronden. Heel erg bedankt dat jullie altijd voor ons klaarstaan en voor jullie rotsvaste vertrouwen in mij. Pap, zonder jou had ik dit jaar geen tijd gehad voor onze heerlijke zomervakantie aan het meer van Annecy. Ontzettend bedankt voor het verzorgen van de lay-out van mijn proefschrift. Anne en Malcolm heel erg bedankt voor de correctie van het Engels.

Lieve Stan en Tijn, mijn reuzenkleuter en grote peuter, wat ben ik trots op jullie! Eindelijk is het zover: mijn 'oude' werk is af. Vanaf nu heb ik weer veel meer tijd om leuke dingen met jullie te doen: gezellig samen kletsen, al jullie vragen zo goed als ik kan beantwoorden, samen spelletjes doen, in de tuin werken, naar het bos gaan, boekjes lezen, knutselen, sommen maken, fietsen, treinen kijken en natuurlijk voetballen. Ik verheug me erop!

Lieve Sander, wat ontzettend fijn dat jij altijd in me bent blijven geloven. Doordat je mij het afgelopen jaar continu gesteund hebt en vaak de zorg voor onze jongens op je hebt genomen tijdens de weekenden en in je vakanties, heb ik mijn proefschrift dit jaar af kunnen ronden. Wat zal er een last van onze schouders vallen en wat zullen we samen gaan genieten van onze vrije tijd!

Curriculum Vitae

Karen Keune was born in Nijmegen, the Netherlands, on September 4, 1979. In 1998 she obtained her Gymnasium diploma at the Stedelijk Gymnasium Nijmegen, after which she went to Groningen to study English Language and Culture at the Rijksuniversiteit Groningen (propaedeutics), and enrolled in the post-propaedeutic programme Information Science at the same university. She specialized in Computational Linguistics and obtained her master in 2003. In december 2003, she started a PhD project within a larger NWO project entitled 'Gesproken Standaardnederlands in Nederland en Vlaanderen: interne variatie en onderlinge verschillen' (dossiernummer 205-41-167), carried out at the Interfaculty Research Unit for Language and Speech (IWTS, now CLSM) of the Radboud University Nijmegen. After giving birth to two sons, she finished her dissertation in 2012. Since October 2011, Karen works as a research scientist at the Psychometric Research Centre of Cito: Institute for Educational Measurement.